Towards Domain-Independent Multi-Lingual-Dialectal Online Social Behavior Modeling

by

Fatimah Alzamzami

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science

uOttawa

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Declaration of Authorship

I hereby certify that this work is entirely my own original work except where otherwise indicated. I am aware of the University of Ottawa regulations concerning plagiarism, including those regarding consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Disclaimer

This thesis uses terms, sentences, or language that are considered foul or offensive by some readers. Owing to the topic studied in this thesis, quoting toxic language is academically justified but I do not endorse the use of these contents of the quotes. Likewise, the quotes do not represent my opinions and I condemn online toxic language.

Abstract

Online social networks (OSNs) have changed the way humans communicate. They seem to have transferred their real-life means of communication and their social behaviors to digital forms on the virtual social media. With this move to the "new world", they have not only adapted some of the already existing forms of communication to fit the new milieu, but they have also adopted new forms of communication provided by OSNs. With communications being shifted increasingly to OSNs, especially after the outbreak of COVID-19 pandemic, the need for tracking and understanding human behaviors online has risen. Also discovering emerging trends and concerns in order to understand the corresponding online social behavior (OSB) that best reflects its offline settings has become a necessity. Further, voicing out concerns and communicating timely trends are not restricted to a single spoken language; Facebook alone reported that two-third of its users speak languages other than English. Besides, the informal and slang nature of conversations and communication has become the new norm on social media platforms which, in turn, has triggered the need to understand foreign languages and even their dialects in order to be able to widely monitor OSB within countries and across the world. This is particularly vital to ensure stability and well-being in societies and to enhance the quality of decision-making in smart cities. Despite all those challenges, we have been able to analyze the users' OSBs, based on the users' textual and visual forms of communication as a first step. This does not involve literal translation, i.e. rendering a text from one language to another without considering the sense of the text, but rather includes examining the geo-cultural contextualization of OSN communications. This is significant as behavior is the outcome of a culture and is manifested through word use (language/dialect in this case). In response, we propose a multimedia framework for modeling domain-independent OSB in different languages and dialects used on OSNs. Unsupervised and supervised learning approaches have been utilized in developing the components of the proposed framework. The first component refers to content-localization based machine translation and is responsible for capturing the multi-lingual multi-dialectal aspect of OSN conversations using AI power. The second component is responsible for modeling textual and visual OSB using machine learning and deep learning algorithms. The third component presents topic modeling and dynamic topic interpretation and is responsible for inferring hidden patterns from a stream of multi-lingual-dialectal data and providing comprehensible interpretations as a step towards facilitating the analysis of the predicted OSB. Further, new datasets have been proposed and constructed to develop and evaluate our proposed AI-models. In addition to our comprehensive experimentations conducted to evaluate the proposed framework, our large-scale analysis of COVID-19 pandemic has reinforced the capability of our proposed solution to recognize concerns and trends, along with reliability to analyze multi-lingual-dialectal OSBs, using real OSN data collected from North America and Middle East regions. This thesis presents a comparative analysis of OSB and data discoveries in Canada, USA, Lebanon and Saudi Arabia.

Acknowledgements

"هَذَا مِنْ فَضْلِ رَبِّيْ"

"This is a favor from my Lord"

I would like to extend my deepest appreciation and gratitude to all those people, without whom I would not have made it through my PhD journey. I would like to express my profound thankfulness and respect to my supportive supervisor professor Abdulmotaleb El Saddik. His guidance, patience, and profound insights have motivated me to persevere this inspiring endeavor. He has been caring and understanding throughout my PhD studies. This thesis would not have been successfully delivered without his genuine support, constant encouragement, and immense knowledge. He has been encouraging for crazy research ideas and is always happy to provide priceless professional and personal advice. It has been an honor to be supervised by such a tremendous mentor.

I would like to express my genuine gratefulness to my beloved family. Thanks to my father who instilled the value of education and knowledge within me from a very young age. Warm thanks go to my mother for her unconditional love. Without her pure love, support and faith in me, I would not have been who I am and where I am today. I would like to express my genuine gratefulness to my kind sisters, supportive brothers, and lovely nieces and nephews for always being there for me. Thank you for being a nice part of my life.

What is the life without family-like friends!. Sana'a, Abdullah, Fedwa, Hani, Ghada, Nuha, Bushra, and Basem, I am short in words to express my sincere gratitude towards your moral support and limitless encouragement you have provided me with during the crucial times in my life. Thank you from the bottom of my heart to you all.

My acknowledgement would be incomplete without thanking my wonderful friends whom I spent memorable times with in MCRLab; Kareem, Hawazin "our spiritual mother", Samah and Rajwa "my all-nighter companions", Majd, Faisal A., Mohammed A., and Majdi. I would like to extend my thanks to my amazing colleagues in MCRLab; Juan, Roberto, Zijian, Haopeng, Monika, Kamran, and Rahatara for not hesitating to give a hand whenever I needed to.

Finally, I would like to express my appreciation to the participants who showed a full commitment throughout the process of dataset construction in this thesis. Special thanks go to Suzanne, Sawsan, Yousef, Mawahib, Raghad, Nooraldeen, Nader, Wisam, Zainab, Isla, Khadijah, Ahlam, Layan, and Sadeel.

Dedication

This thesis is lovingly dedicated to the soul of my beloved mother for her pure and unconditional love. Even though you are no longer with me, I can still feel your love and prayers guiding me. No words or feeling can ever describe how much I am grateful for everything you had given me.

May your soul rest in heaven...

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

It is needless to say that online social networks (OSNs) have transformed communication from just simple phone text messages to today's instant and temporary stories, and the dynamics of social communication have significantly changed in the past years. The number of OSN users has mounted from 970M in the last decade to 5B users in 2023. Those users actively use OSN platforms worldwide as an everyday medium for communication and content sharing; this figure constitutes more than 60% of the world's population. It is worth noting that mobiles have contributed a great deal in the rise of the number of social users, as $\approx$ 99% access OSNs from their smart phones.

Social media has provided people with a wide range of opportunities to communicate on a large scale and to constantly share visual contents, respond to textual comments, or update statuses -be it a permanent post on a timeline or a temporary story. In response to shared posts, users reply or comment soon enough, which creates a sense of a real-life form of communication. That being said, social media has not only created an indispensable need to share a content and be shared, but has also become an integral part of people's daily activities as evident by statistics[1] (Figures 1.1 and 1.2).

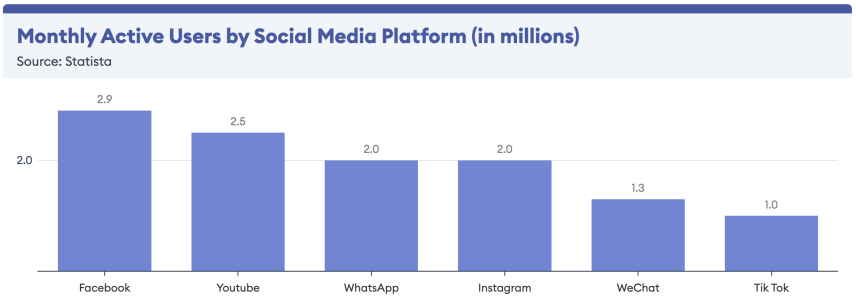


Figure 1.1: **Social media statistics - Monthly actives users - 2023, according to Forbes[1].**

OSN platforms can be considered ideal forums that reveal a great deal about people's interaction and reaction to both hard news and soft news such as elections, pandemics,

[1] https://https://www.forbes.com/advisor/business/social-media-statistics

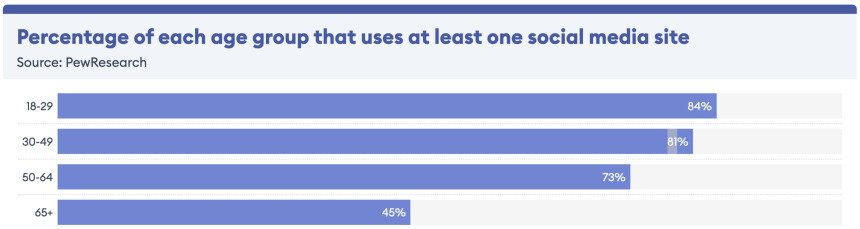


Figure 1.2: **Social media statistics - Age group - 2023, according to Forbes[1].**

weather extremes, soccer games, food or fashion. A wide array of people from different places of the globe share their views and perspectives regarding most of the stories that have become globalized due to interconnectedness. A perfect example of social media reach is seen in the way the world communicated during COVID-19 pandemic. The communications during that period shifted to online social networks (OSNs) where people would not only read and share timely information about warnings, announcements and statistics, but also voice out their reactions regarding the social measures taken to halt the spread of this infectious disease.

Humans communicate through utilizing different types of language, and they reflect their thoughts, opinions, feeling, and behavior through the use of words, tone, pace, facial expressions, body language, etc. Ever since people turned to social media to communicate their everyday details, they have transferred with them their means of communication and their social behaviors to a digital form. Those means, which were either introduced by the OSN platforms or have been adapted by users to fit the new medium, include but not restricted to verbal-written communication, pictorial communication, and video communication.

Understanding and analyzing human online social behaviors through the users' textual and visual forms of communication have become feasible. With over 50% of the world population creating daily personal contents on social media, OSNs definitely host human digital replicas that preserve users' identities and all communications expressed in digital forms. From such forms we can detect whether the users are happy about a movie or dissatisfied with a product, for example. Research efforts have been exerted to interpret and analyze different types of online social communication. In this context, substantial progress has been made in examining behaviors, sentiments and emotions- positive and negative- of social users in the past few years [8,17,19,35,214,218]. Special attention has been placed on violence and hate behaviors through implementing intelligent systems that detect various phenomena of toxicity on social media for the purpose of maintaining social stability [13,17, 33,151]. In this context, recognizing sentiments, emotions and toxic behavior on OSNs has served as a valuable tool to monitor public health during pandemics, for example [17]. In addition, detecting and examining other sentiments and feelings such as sarcasm and humor on OSNs have been an active research arena in social media analysis to understand people's true opinions [70–72]. Lately, studying the behaviors of optimism and pessimism has been utilized to reduce the negative attitude on social media [11,12]. However, analyzing such behaviors on OSNs is not applicable to many languages especially those that branch into different regional dialects. This is due to the limitation and insufficiency of their resources

required to build the necessary models. With the world having been shifted to online communications, this limitation triggers an urgent necessity to address this gap and discuss adequate and flexible solutions that facilitate the understanding of online social behaviors in smart cities, within and across geographical regions, with the goal of improving the authority decision-making and hence, citizens' QoL.

1.1 Research Motivation

The concerns of the domain-specific online social behavior (OSB) modeling shed light on the importance of generalizing the social behavior learning independently of any other domains. Current OSB models trained using domain-specific datasets, usually come with the concern that they do not generalize to other domains [17] since they latch to the information of the domain they have learned from [17, 228]. Besides, existing publicly available datasets have been either collected with pre-defined topics and keywords or noisily annotated [104, 222]. Conversely, general knowledge of models is more effective in terms of speed and cost. An advantage of the general knowledge of models is that it would help with speeding up the learning process of specific domains or the process of domain adaptation. Another advantage for building general domain-independent models is that a reasonable performance in terms of resources and training can be obtained at a lower cost than building a model for every domain. Those two features have motivated us to take the first step towards solving the domain-free problem and creating a domain-independent multimedia online social behavior dataset. Our proposed dataset follows high quality protocols and techniques to address the purpose of this thesis. Note that -in this thesis- we study two types of behaviors: sentiment and expression toxicity.

People use different languages, native or/and non-native, to communicate and share information on different social media platforms. According to Facebook, two-third of its users use languages other than English, of which Arabic, Spanish, and French are among the top languages used on OSNs. In order to accurately and automatically analyze online communications expressed in different languages and dialects, intelligent models need to be built with the capability of understanding the communication expressed in those languages and their dialects in order to better preserve the meaning in their contexts. The semantic of expressions may vary across languages and dialects due to several factors, such as language development, the influence of other languages on the native language and the cultural influences [106,210,229]. There is diversity in the dialect people use to express opinions and emotions, depending on socio-cultural contexts. Therefore, semantic orientation varies with cultural differences. For instance, the Arabs, who constitute 11M active users on Twitter and create over 27.4M tweets everyday [10], do not all use Standard Arabic on social media; instead, they use dialectal Arabic common in social communications especially in informal conversations. However, dialectal Arabic, which differs significantly from Standard Arabic, varies from one region to another [237] in terms of phonology, morphology, syntax, and lexicons. So far the existing works that have utilized Standard Arabic resources for online social behavior analysis have not been able to perform well on dialectal Arabic data in the context of online social analysis [185] due to the fact that the resources for dialectal Arabic

are scarce. Also, the research trend tends to solve a specific problem in a specific language or dialect [2,10,30,71,89,110,150], instead of utilizing the already existing resources to analyze OSB in low-resource languages and dialects. This can be seen in the huge discrepancy of resources between languages, where very few have a high-resource status while many others are low-resourced. English, for instance, is a high-resourced language while $> 70\%$ of OSN users speak languages other than English, of which Arabic, Spanish, and French are among the top used languages on social media yet they are considered low-resourced languages. This research trend limits its current systems to generalize to other different domains and languages [17,19,228]. Moreover, the expensive cost (i.e. in terms of HW/SW requirements, time, efforts, cost, and human labor) for building a resource for each language and dialect has definitely contributed to the under-resourced status of existing multi-lingual-dialectal systems. Researchers are obligated to address the shortcomings of this current practice and study reliable yet inexpensive solutions that minimizes the dependency of domains and languages/dialects in modeling online social behaviors. This thesis addresses this limitation and proposes a framework that is specifically designed to handle the multi-lingual-dialectal aspect of online social behavior analysis on social media. The proposed framework is cost-effective and aims to minimize the language and domain dependency issue to analyze various online social behaviors on online social media.

The immense amount of social media publicly published data from various domains on different topics and trends triggers an urgent need for compiling such massive data for the purpose of exploring the underlying topics, concerns and trends in a timely manner, and facilitating the analysis of online social behaviors. However, it is nearly impossible to achieve this manually. Consequently, machine intelligence is inevitably necessary to automate the monitoring of online stream of conversations and discover hidden topics and trends, as well as coherently interpret them. The unsupervised learning methods make it possible to achieve this goal fast and without prior human knowledge involved. However, these methods are sensitive to data noises, which is a normal phenomenon in OSN data. It is well-known that social media data suffers from noises such as uncontrolled visual content, misspellings, slang usage, and the intense use of abbreviations as a result of the limited writing-space capacities. This challenge requires a careful handling of the OSN data for the intelligent models to learn and perform well. In response, we propose a methodology for data exploration and dynamic interpretation, as described in this thesis.

1.2 Research Objective

The main objective of this thesis is to leverage the multi-lingual multi-dialectal contents of social media in modeling domain-independent online social behaviors. To effectively analyze multi-lingual-dialectal online social behaviors, a model capable of handling the local cultural expressions of different languages and dialects has to be created. In order to achieve this, a high quality mechanism is needed to handle not only the unique nature of each language but also the different dialects within the same language, in order to guarantee the reliability of a developed model that adequately understands the expressions in context, and then accurately delivers an analysis of the underlying OSB of interest.

Previous studies on multi-lingual OSB analysis have some limitations or flaws. They adopt classical translation to bridge the gap of resource insufficiency of low-resourced languages. Moreover, they totally overlook the dialects within each language, and in so doing, they fail to consider the local culture of the language or dialect as a key factor. All of this leads to incomprehensible target text by native speakers specially when dealing with collocations, idioms, proverbs, preposition usages. As a result, the reflection of the underlying online reactions and social behaviors (e.g. sentiment or emotions) is misinterpreted and misconceived. In response, this thesis aims to take into consideration and pay special attention to the local culture of a language that resides within the wealth of introspective OSN data for the purpose of adhering to the real meaning of the communicated expressions, and hence be able to automatically expand the analysis of various online social behaviors across different languages and dialects. To achieve this, our research is divided into four main stages. First, we develop content-localization based machine translation models to localize OSN contents from a source language to the specific culture of the target language or dialect. The objective of this stage is to make it possible to exploit resources of both low-resourced and high-resourced languages and dialects in order to be able to reliably analyze online social behaviors in different languages and dialects on online social media. Second, we focus on designing and developing a domain-independent models to track the online social behaviors in different domains across different languages and dialects. The objective of this stage is to minimize the dependency of domains in order to be able to generalize our model to solve OSB analysis in various domains. Note that this thesis considers two types of social behaviors as a case study: sentiment and toxic speech. In the third stage, we utilize images as a supplementary language-independent approach where a universal language of visual images and emotions are used to model online social behavior. The objective of this stage is to compensate for missing or insufficiency of language resources since images are language-independent medium for expression. Also, it provides a visual view for users' reactions in which it enriches the online social behavior analysis since images can convey a lot of information in a single glance. In the final stage, we develop a model for OSNs real-time data exploration and coherent dynamic interpretation to be used as a complementary tool to facilitate the analysis of online social behaviors. The objective of this model is to automatically infer insights and trends from OSN data and dynamically generate comprehensive interpretations of those insights and trends.

This thesis tackles the complexities of analyzing OSBs in different domains and across different languages/dialects. It achieves this by outlining a specific process consisting of several requirements as follows:

- Support the textual and visual domain-independent prediction and analysis of online social behaviors. This will support generalizing OSB models to different domains. This requirement is addressed in chapter 4 and 6.

- Support the multi-lingual-dialectal predictions of online social behaviors. This will support the content localization of OSN messages from a source language to a target language and dialect. This will be used to create missing and/or insufficient resources for low-resourced languages. The visual analysis also supports the cases where data

resources of a language is missing or insufficient. This requirement is addressed in chapter 5 and 6.

- Detect hidden insights and trends from OSN data streams and automatically provide dynamic interpretations for those inferred insights and trends which will be used to utilize the analysis of online social behavior. This requirement is addressed in Chapter 7.

We summarize the key research questions that this thesis addresses as follows:

- How to effectively analyze online social behaviors in different domains of different languages and dialects?
- Do differences in dialects affect the analysis of OSB?
- Can we adapt general online social behavior models to domain-specific problems?
- How to infer high level insights and trends from OSN data stream? And can the inferred insights and trends be automatically and coherently interpreted in different languages/dialects?

1.3 Research Challenges

The extensive flow of information streams could efficiently be handled using AI powers to continuously keep track of public current states since it is nearly impossible to manually monitor huge loads of online data flow. While developing our domain-independent multi-lingual-dialectal online social behavior model, we encountered the following challenges:

- **OSN Open Platforms:**

  Social media platforms like Twitter and Facebook are open platforms (i.e. domain-free platforms) where people can talk about anything and everything varying from expressing personal opinions in politics to chanting for favorite sport team and reviewing products and movies. In the literature, researchers tend to propose customized solutions for specific domains [110] such as customized sentiment model for soccer fans [7] or sentiment model for movie reviewers [197]. While this approach works well on problems of similar specific domains, it has been proven that it does not generalize well to other domains since they latch to the information of the domain they have learned from [17, 19, 228]. Indeed, it is resource-expensive and time-consuming to build a model for every domain especially with the continuous content generation on OSNs. COVID-19 pandemic is a great example that shows that those domain-specific models do not work well on a newly emerged data of different domain. This thesis addresses this issue and studies the domain-independent online social behavior analysis as an attempt to minimize the dependency of domains to solve online social behavior analysis.

- **Language Diversity and Informal Communication:**

  The more the social communities expand, the more challenging it is to manage the information streams. There are numerous user-generated contents on OSNs that have not been exploited sufficiently, especially for multi-lingual-dialectal purposes. Classical translation across languages has proven to be ineffective in analyzing online social behavior [175] since it ignores the context of messages. As mentioned above, creating a data resource for every problem is not a cost-effective solution and that is seen in the scarce data resources of low-resourced languages like Arabic. In this thesis, we address this gap and propose a content-localization based machine translation approach in order to allow for the exploitation of resources in low and high resourced languages and dialects while preserving the semantics and contexts of contents, and hence generate a relevant analysis for online social behavior in the language of interest.

- **Data Noise and Information Overload**

  Online social networks encourage unstructured data format that does not follow grammar conventions. It is well-known that social media data suffers from noises such as misspellings and the intense use of abbreviations due to the limited writing-space capacities. Images in social media are freely shared. Thus, it is extremely complex to find the relationship between this diverse data and its online social behavior (e.g. sentiment) orientation, which makes the semantic gap problem very serious [125]. Such informal and noisy nature of data makes social media analysis more challenging compared to the analysis done on a data of structured format. The data noise combined with the extensive size of OSN data bring difficulties to the modeling of online social behaviors. This thesis tackles this issue by utilizing high-quality techniques and approaches that reduce the sensitivity to data noises with the goal to overcome the mentioned challenges.

- **Data Imbalance**

  It is a common phenomenon to have data imbalance on online social behavior datasets collected through OSNs [26]. Literature [226] suggests that having balanced dataset would improve the learning process. However, it is too expensive and time consuming to balance the data while preserving the natural distribution to avoid biases. This thesis responses to the data imbalance challenge and proposes solutions to overcome its limitations with the objective of improving the learning of online social behaviors.

- **Dataset Construction**

  Actually, building a dataset from scratch is very costly and time-consuming. During the process of creating our datasets, we have observed the following challenges: (1) data collection requires extensive efforts to search and match the required data, including data fetching, retrieval, cleaning, and filtering of the data. (2) data annotation and translation need domain experts. It is very hard to have experts agree to work on a large volume of data, and even if they do, the cost in terms of time and expenses is very high, let alone the tediousness of the task.

1.4 Thesis Contributions

We summarize the contributions of this thesis as follows:

1. Propose a multi-lingual-dialectal framework for domain-independent online social behavior analysis.
2. Design and construct content-localization based translation dataset designed to translate OSN conversations (i.e. tweets) from multiple languages (i.e. English, French, and Spanish) to Arabic multi dialects (i.e. Gulf, Levantine/Shami, Iraqi, and Yemeni).
3. Develop and implement content-localization based neural machine translation models designed to translate OSN conversations (i.e. tweets) between multiple languages (i.e. English, French, and Spanish) and Arabic multi dialects (i.e. Gulf, Levantine/Shami, Iraqi, and Yemeni).
4. Design and develop intelligent multimedia models for domain independent online social behavior analysis in multiple languages/dialects.
   (a) Design and construct a domain-free multimedia dataset for sentiment analysis.
   (b) Develop multi-lingual-dialectal textual classifiers for sentiment and toxic behavior analysis using tweet texts.
   (c) Develop a multi-modality visual classifier for visual sentiment behavior analysis using tweet images.
5. Design and implement an OSN-specific multi-lingual model for topic modeling and dynamic topic interpretation using OSN conversations (i.e. tweets) as a step towards facilitating the online social behavior analysis.

1.5 Thesis Organization

The rest of this thesis is organized as follows: Chapter 2 reviews the background related to the work presented in this thesis. Chapter 3 presents an overview of our proposed framework for modeling domain-independent online social behavior in different languages and dialects. The design and development of the domain-independent online social behaviors is described in Chapters 4 , while a comprehensive description of the design and development of multi-lingual-dialectal online social behaviors is detailed in Chapter 5 and 6. Chapter 7 presents comprehensive details of topic modeling and dynamic topic interpretation followed by the conclusion and future directions in Chapter 8.

1.6 Scholarly Achievements

- **Research Resulted in Refereed Journals**

1. **Fatimah Alzamzami**, Abdolutaleb El Saddik, Content-Localization based System for Analyzing Sentiment and Hate Behaviors in Low-Resource Dialectal Arabic: English to Levantine and Gulf. 2023 (submitted)
2. **Fatimah Alzamzami**, Abdolutaleb El Saddik, OSN-MDAD: Machine Translation Dataset for Arabic Multi-Dialectal Conversations on Online Social Media. 2023 (submitted)
3. **Fatimah Alzamzami**, Abdolutaleb El Saddik, Transformer-based Fusion Approach for Multimodal Visual Sentiment Recognition using Tweets in the Wild. IEEE Access, 2023
4. **Fatimah Alzamzami** and Abdulmotaleb El Saddik. Monitoring Cyber SentiHate Social Behavior during COVID-19 Pandemic in North America. IEEE Access, 2021
5. **Fatimah Alzamzami**, Mohamad Hoda, and Abdulmotaleb El Saddik. Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation. IEEE Access, 2020
6. **Fatimah Alzamzami**, Mukesh Saini, and Abdulmotaleb El Saddik. DST: Days Spent Together using Soft Sensory Information on OSNs: a Case Study on Facebook. Soft Computing, 21(15):42274238, 2017

- **Research Resulted in Refereed Conferences**

7. **Fatimah Alzamzami**, Abdulmotaleb El Saddik. Content-Localization based Neural Machine Translation Approach for Sentiment and Hate Speech Analysis in Arabic dialects: Spanish/French to Levantive/Gulf Arabic Dataset. 2023 (submitted)
8. Rana Abaalkhail, **Fatimah Alzamzami**, Samah Aloufi, Rajwa Alharthi, and Abdulmotaleb El Saddik. Affectional Ontology and Multimedia Dataset for Sentiment Analysis. In International Conference on Smart Multimedia, pages 1528. Springer, 2018
9. Samah Aloufi, **Fatimah Alzamzami**, Mohamad Hoda, and Abdulmotaleb El Saddik. Soccer Fans Sentiment through the Eye of Big Data: The UEFA Champions League as a Case Study. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pages 244250. IEEE, 2018

Chapter 2

Related Work

2.1 Background

2.1.1 Online Social Behavior (OSB)

In real- life settings, people's communicative interaction is governed by the means of symbolic-based cultural systems, part of which are the shared rules of language and other verbal and non-verbal collective symbolic systems. [192]. Communication could "include the information that people acquire, through inference, from all kinds of interactors participating in material medium," [192]. The entire set of communications taking place during an interval of time with reference to a group of people can define what it is called a behavioral system [191]. Accordingly, "a behavioral system is composed of all people-artifacts, people-people, people-extern, artifact-artifact, and artifact-extern interactions relating to the members of a specific household or community or society" - Michael Schiffer [192]. An operation of behavioral system could be looked at as a repetition of its constituent interactions [192]. In this context, Schiffer [192] claims that communication and behavior are related to each other, and both consist of people-people, people-artifact interactions. Similarly, Hannema [90] claims that "when we interact with others, anything we do communicates. Behavior is communication and communication is behavior". Further, Schiffer distinguishes between human behavior seen in stimulus and response and the causes of behavior such as goal and intention.

With the evolvement of social media from information exchange to virtual meeting places, people seemed to have moved their real-life communications to digital format by using social media platforms, where individuals interact with each other and with virtual entities within same or different geo-locations at different times. As a result, new definitions of communication and behavior have been introduced. That being the case, we can observe a repetition of interactions that share common characteristics (e.g. positive reviews on a new product). According to Schifer, a human behavior is formed of its constituent interactions that share common characteristics. Since the focus of this work is on the communication on social media, we refer to the behavior as online social behavior (OSB). Online social behavior (OSB), or online interpersonal behavior, can be seen in sentimental

and emotional feelings (i.e. positive or toxic sentiment for instance) expressed towards an event or a situation. The events could be soft news stories like a soccer game, movies, and fashion or hard news stories like extreme weather, a coup d'état, an epidemic or a pandemic.

2.1.2 Online Communication Language

Online communication language refers to the communicating mediums used among OSN users to interact with each other. These mediums take the forms of spoken languages, pictorial and iconic interactions, mouse-click style interactions, audio conversations, and video communications. The spoken language-based communications are conveyed through verbal, textual and/ or pictorial interactions. Like in real-life settings, language production requires that the person communicating have prior knowledge of the language(s) of interest in order to have successful back-and-fourth communications. Spoken languages all around the world extend to various dialects and even sub-dialects of 'the one' language; hence, they are not limited solely to the official languages within geographical regions, which makes it sometimes hard to understand a conveyed message. Nevertheless, pictorial-based communications enable people to universally communicate across languages and dialects without the need for prior understanding of a specific speech; hence, unlike verbal language that introduces language-dependent interactions, images communicate information through their visual elements, such as colors, scenes, shapes, and faces, etc. An image of a smiling face, for instance, can denote happiness and positivity. This sentiment can be expressed pictorially, without the need for a verbal expression. Static/animated images, emoticons, and emojis are all examples of pictorial communication over online social networks. The availability of such different types of online communications give rise to the need for utilizing their unique characteristics in the modeling of online social behavior on OSNs.

2.1.3 General Knowledge vs Specialized Knowledge

Knowledge, in general falls under two types, general knowledge and specialized knowledge. General knowledge can be defined as a broad range of knowledge that covers a variety of topics in different domains, whereas specialized knowledge is more in-depth knowledge about a particular topic or domain. While specialized knowledge is more in-depth than general knowledge and can be helpful for solving complex problems in specific domains [227], it is often less transferable than general knowledge and somehow limits its applicability to solve a wider range of problems. Conversely, general knowledge is more diversified and more capable of understanding a wide range of topics; hence, it is more capable of providing flexible solutions to challenges in various fields and domains. [47, 83]. Having a strong foundation in general knowledge can serve as a valuable asset in a broader range of fields and provide learners with a broad base that helps them adapt to different areas of interest efficiently [47]; thus, one can remark that general knowledge sets a solid ground to specialized knowledge [47]. By the same token, the global education system implements

the progressive specialization approach [47, 135, 213]. Progressive specialization approach is based on the idea that learners should start with "a broad general education", a broad foundation of knowledge, and then gradually specialize as they move through the different levels of education (i.e. college and beyond). The shift from general to specialized education is a natural progression as learners progress through their education. In high school, for instance, students typically take a variety of courses in different subjects, such as English, math, science, and history. In college, they begin to narrow their focus on the field of their interest, and in graduate school, they specialize even further and gain the skills and knowledge they need to be successful in their field. The progressive specialization is widely accepted as a sound approach to education as it allows students to be more versatile and adaptable in the workforce [213]. The effectiveness of the progressive specialization theory in education highlights the importance of general knowledge throughout learning process especially in the early stages of learning [135, 213].

Based on what has been mentioned, the ability of general knowledge to provide a solid foundation to effective understanding and solving more specialized problems is what motivated us to adopt the general knowledge approach and utilize its benefits for machine-learning-based modeling for online social behaviors. Note that we refer to general knowledge as domain-free knowledge or domain-independent knowledge interchangeably. In this thesis, we define domains as specific areas of topics where words and expressions are used in a particular way. Domains could be broad such as "politics" or "healthcare", or narrow such as "movie reviews" or "product reviews".

2.1.4 Traditional Machine Learning Algorithms

2.1.4.1 Gradient Boosting Machines

Gradient boosting is one type of ensemble learning. Unlike classic learning approach, ensemble learning approach combine a set of weak learners to construct one strong learner [116]. In contrast to the bagging technique where the models are made independently, the models in the ensemble boosting technique are made sequentially by iteratively minimizing the error of earlier learnt models [50]. It learns a predictive model by combining the M additive tree models $(f_0, f_1, f_2, ..., f_M)$ to predict the results (Eq.2.1).

$$f(x) = \sum_{m=0}^{M} f_m(x) \tag{2.1}$$

The tree ensemble model is optimized by reducing the expected generalization error L according to Eq.2.2:

$$L = \sum_{i}^{n} (y_i - \hat{y_i})^2 \tag{2.2}$$

L is a loss function that measures the delta loss between the target y_i and the prediction $\hat{y_i}$ of a data point.

There are three fundamental reasons listed by Dietterich [60] to use an ensemble-based methods:

- **Statistical**: combining and averaging multiple learners provide a better generalization on the learning of data which in turns, reduce the risk of choosing inadequate classifiers.
- **Computational**: during learning, it is computationally difficult for an algorithm to search for a local optima in order to learn the best representation (i.e. decision boundaries) of the data. Neural network algorithms, for instance, utilize gradient descent to minimize the loss function during the training to learn the best model. In this case, there is only one starting point for the local search. With the ensemble algorithms, we have an advantage of having multiple starting points for the local search. This may provide a better approximation of the true function (i.e. decision boundary) than an individual classier does.
- **Representational**: there are cases where a single classifier is not able to learn a decision boundary that separate different classes, or the decision boundary is very complex. Here comes the advantage of the ensemble-based learning where it provides different decision boundaries learnt from different classifiers.

For the reasons mentioned earlier, we believe that using ensemble gradient boosting helps to increase the robustness of classifiers while decreasing their variances and biases. The nature of the boosting technique could decrease errors as it reduces the failures of individual classifiers while optimising their advantages at the same time. Hence, a more reliable model could be produced. In this work, we utilize two powerful gradient boosting algorithms: Extreme Gradient Boosting (XGB) and Light Gradient Boosting Machine (LGBM). They are the state-of-the-art algorithms from the gradient boosting family. XGB was introduced in 2016 [50] and LGBM was introduced by Microsoft in 2017 [111]. We train XGB and LGBM classifiers to give the sentiment class (i.e. positive, neutral, negative) based on the five types of features explained later on in Section 4.3.1.1.

2.1.4.2 Extreme Gradient Boosting (XGB)

XGB [50] is an ensemble tree-based method that implements a gradient boosting machine learning framework for regression and classification problems. XGB grows trees using level-wise algorithms. It differs from RF in the way it grows, orders, and combines the results. XGB uses different algorithms for splits finding. Exact Greedy and Approximate algorithms were introduced first in [50]. Histogram-based algorithm was then proposed to be used for the splits finding after LGBT algorithm was invented. When histogram is used, trees grow in leaf-wise manner.

The method works by bucketing features values into group of bins to construct features histogram. The splitting is performed on the bins instead of on the features. The bucket bins are constructed before each tree is built, hence, it speeds up the training which in

turns reduces the computation complexity. In this work, we use the histogram method for deciding the best split. During the parameters turning, we found out that the three algorithms yield similar results. Therefore, we decided to go with histogram method since it takes faster training time on large sparse datasets, than the other splitting algorithms.

The decision to make a split is based the loss value that the split produces. The split will happen if the loss value exceeds a certain threshold, otherwise, the split will be ignored. This shows the advantage of leaf-wise gradient boosting methods over the RFs in reducing the number of splits while keeping the quality of the splits.

Furthermore, XGB uses sparsity-aware split algorithm that works on sparse vectorized textual data (i.e. the case in our work). When computing the split, the sparsity-aware split algorithm proposes to ignore the zero features, and then allocates all the data with zero values to the side of the split that reduces the loss the most.

2.1.4.3 Light Gradient Boosting Machine (LGBM)

LGBT [111] is another gradient boosting algorithm that uses a leaf-wise algorithm to grow trees vertically.

A leaf that reduces the loss the most is chosen to split and grow the tree. LGBM uses histogram-based method to find best splits candidates.

LGBM is able to deal with uneven data distributions. Its natural design allows it to deal with data imbalance through the Gradient-based One-Side Sampling (GOSS) technique. GOSS is a sampling algorithm that indicates the importance of data instances. Its main function is to concentrate on data samples with larger gradients and ignore the data with small gradients. The assumption is that the data with small gradients have lower errors; thus they are already well trained. Therefore, GOSS proposed to ignore these less-informative data points and use the rest to compute the information gain when finding the best splits. However, this will result in a bias problem towards the sample with larger gradients, and will change the original distribution of the data. To solve this issue, GOSS performs a random sampling on the data with small gradients while keeping all the samples with large gradients. Because the sample would still be biased towards the data with large gradients, GOSS increases the weights (i.e. adding a constant multiplier) of the data instances with small gradients when computing the information gain.

In addition, LGBM uses Exclusive Feature Bundling algorithm to handle sparsity in datasets. It combines mutually exclusive features in a nearly lossless way resulting in reducing the number of features while keeping the most informative ones.

2.1.4.4 Support Vector Machine (SVM)

SVM algorithm has shown a robust performance in text classification [223]. The goal of SVM is to select a hyperplane that maximizes the margin between the closest instances of the two classes.

Sentiment analysis in this work is a multi-class classification. SVM is by default a binary-class classifier. We follow "one-vs-all" approach to solve the multi-class classification problem using SVM. Note that we attempt to use linear kernel in our experiments. The initial experiments with non-linear SVM have shown a decreased learning performance.

2.1.4.5 Logistic Regression

Logistic regression is considered one of the best discriminative models. It learns the posterior class probability directly from the training data. Its goal is to find a decision boundaries between classes in the feature space. The posterior probability, in binary classification, is given by applying the *sigmoid* function on a linear combination of given inputs. Logistic regression can be generalized to work on multi-class classification problems by using $softmax$ function to derive the posterior probabilities by normalizing a given feature vector to probability values values between $[0, 1]$.

2.1.4.6 Multinomial Naïve Bays

Naïve Bayes is a probabilistic classifier. Though it is called naive, it performs well in text categorization [172]. The core of the algorithm is based on Bayes theorem. The Multinomial Naïve Bayes (MNB) classifier assumes multinomial distribution so that it can be used with discrete features like words counts in text classification. In our work, we attempt to use the Multinomial Naïve Bayes classifier for our three-class sentiment analysis problem.

2.1.4.7 Random Forest (RF)

Random Forest (RF) is an ensemble tree-based classification algorithm. It uses bagging techniques in which trees are fully grown to their maximum extent. The trees, in RF, are trained independently using a random sample of data. Every tree in RF is generated based on bootstrapped training instances and a random set of features. Each learnt tree is a weak learner. By combining all the weak learners, we have one final strong model. The overall prediction of the RF is computed based on the majority votes from all the individual weak learners (i.e. individual trees). RF has shown a robust performance to noise and overfitting problems that would affect a single decision tree [180]. Moreover, RF can efficiently handle large size of data and is inherently suited for multi-class problems.

2.1.5 Deep Learning Algorithms

2.1.5.1 Long Short Term Memory (LSTM)

LSTM is an extension of standard Recurrent Neural Networks (RNN) which is capable of learning long-term dependencies between words in sequences. LSTM was desinged to overcome the gradient vanishing issue that RNNs suffer from. It was designed with internal

mechanism (i.e. gates) that regulates the flow of information. Its architecture consists of three gates (i.e. input gate, forget gate, and output gate) to decide how much information should flow in (i.e. to remember) and out (i.e. to forget) at the current time step. LSTM models process a sentence word by word and assume that each current state depends only on its previous one. LSTMs process a sequence of words in a forward direction where bidirectional LSTMs (biLSTMs) process the textual sequences in both backward and forward directions. This mechanism allows for more information to be available for the network to improve word contextualization. In this thesis, we train an LSTM model using GLOVE embeddings on both sentiment and hate classification tasks for evaluation purposes.

2.1.5.2 Convolutional Neural Networks (CNN)

CNN is a feed-forward neural network that is biologically-inspired variants of multilayer perceptions. It tends to recognize textual patterns directly from texts with minimum preprocessing applied before feeding sequence of words to the network. A CNN's hidden layer consists of a convolutional layer, pooling layer, and fully connected layer. CNNs strengthen their power from the convolutional layers that are stacked on top of each other with each one capable of extracting unique patterns independently of prior knowledge or human effort. The patterns could be expressions of multiple sizes (i.e. 2, 3, or 4 adjacent words). In this thesis, we build a CNN model with GLOVE embeddings as input features for sentiment and hate classification tasks. We do this step for evaluation purposes.

2.2 Online Social Behavior (OSB) Analysis: State-of-the-Art

2.2.1 Domain-Independent OSB and Existing Datasets

There exists a number of online social behaviors that have been studied in the literature however most of them follow on the domain-dependent approach. Many researchers have focused on domain-specific sentiment analysis [7, 63, 108, 152, 178, 179]. Studies on product reviews and political voting forecasts are examples on domain-specific sentiment analysis [125, 152]. This extends to the methodology that previous studies adopted to collect their datasets. Authors in [6] used emotion keywords to collect tweets while others [8] used domain-related keywords (e.g. event-related or topic-related) and hashtags and smiley emojis [59] to build their dataset. Similar to sentiment behavior, recent trends in toxic speech related researches have focused on specific targets such as racism or aggression [33, 76, 171, 228]. Accordingly, the datasets have also been featured according to the targeted focuses. Authors in [236] studies the offensive language in tweets using a domain-specific keyword based dataset while Waseem and Havoy targeted racism and sexism to analyze hate speech on online social media [224].

A closer look at the existing publicly available sentiment datasets reveals a number of limitations that restrict our purpose for this thesis. One of the major limitations in some of the existing datasets is the focus on the polarity or the valence of the tweets, thus ignoring the neutral sentiment class [63, 152, 204]. For example, Go et al. in [82] constructed a large Twitter training dataset with 1.6 million tweets using positive and negative emoticons as noisy labels. Authors in [170] used similar labeling approach for positive and negative sentiment however they added a neutral class using 44 newspaper and magazine accounts. T4SA [215] is a text-image sentiment dataset where annotations of images were given based on the sentiment of the corresponding text. Automatic or noisy dataset annotation approach [27, 104] compromises the knowledge quality of the learning process. Dmitry et al. [59] utilized 50 hashtags and 15 smileys as labels for sentiment and no-sentiment categories. For the sentiment category, two human-judges manually labeled the 50 hashtags and used them to collect tweets and construct a hashtag-based sentiment dataset. Ten Amazon Mechanical Turk workers labeled 15 smileys with mood states that were used to collect tweets and construct a smiley-based sentiment dataset. For the no-sentiment category, they randomly selected tweets with no hashtags or smileys. Besides the limitation of being restricted to sentiment and no-sentiment labels, relying on hashtags and mood smileys might not truly reflect the sentiment of the tweet. Moreover, tweets in some of the existing datasets were retrieved and labeled with respect to trending topics, specific events or products. For instance, in [82] test data, they manually labeled a set of 177 negative tweets and 182 positive tweets that were collected using queries related to specific companies, events, locations, music, movies, people and products. Sanders dataset [1], the Dialogue Earth Twitter Corpus [2] and the Health Care Reform datase [203] are other domain specific datasets that are publicly available. Tweets included in these datasets are manually labeled for sentiment with respect to specific topics. For example, Apple, Google, Microsoft and Twitter in the Sanders dataset, weather and gas prices in the Dialogue Earth Twitter Corpus, and 8 targets including: Health care, Reform, Obama, Democrats, Republicans, Tea party, Conservatives, Liberals, and Stupak in the Health Care Reform dataset. Even SemEval [182], a well-known dataset that is widely used to evaluate sentiment analysis methods, is constructed based on 200 English trending topics and popular events. MVSA [161] is a text-image sentiment dataset that was constructed based on a set of emotional keywords. In addition to the above limitations, the annotation methodology for most of the manually annotated datasets have some ambiguities. Authors in [82, 203] did not report a detailed description of the data collection and annotation procedures such as the annotators' selection criteria, the number of annotators, their demographic information, and the agreement among them. Moreover, training and evaluation have been conducted on small datasets [40,63,204] and many works have not handled the OSNs cultural language such as iconic emotions (i.e. emojis and emoticons) [27, 79, 152]. It has been repeatedly reported that training deep neural networks using large datasets yields better results than the training using small datasets [64]. Also, iconic emotions contain sentimental clues that would greatly contribute in sentiment learning [19, 99]. In addition to the small data size limitation, most of the exisitng sentiment datasets are either image-based only [40, 233] or

[1] http://www.sananalytics.com/lab

[2] www.dialogueearth.org

textual-based only [7, 59, 108, 152, 170, 179, 182] but did not consider both texts and images while constructing the datasets. Even though cross-domain sentiment learning has been addressed, many studies focus on adapting sub-domain to one another while considering the main domain to be the same [63, 153, 178]. This approach would not generalize well on various domains since those models will latch on to domain-specific information [228].

Similar limitations were found in the existing toxic speech behavior datasets where these datasets have focused on studying one aspect of the multifaceted toxic language behavior. Authors in [236] used keywords such as "gun control, conservatives, liberals" and constructions such "she is, they are"- with an assumption that these keywords are often included in offensive messages- to build an offensive dataset. A hate speech dataset [33] was constructed based on women or immigrants as targets to construct this dataset where abusive and spam aspects were considered by Antigoni [76] in building an abusive dataset. Authors in [189] targeted the problem of aggression and misogynistic identification for three languages on social media. However, iconic emotions (i.e. emojis and punctuation-based emoticons) were ignored. OSNs specific feature such as exclamations marks, words with repetitive characters (e.g looool) were ignored even though they contain strong sentimental insights [19]. Another limitation of the existing dataset is that some existing datasets are small in size. Hind et al. [4] proposed using domain-specific word embedding to detect white supremacist hate speech if it existed on social media. The evaluation was conducted on balanced dataset collected from Twitter and Stormfront forum. However, the size of the proposed training set is quiet small (i.e. 4588 messages) to be able to build a robust classifier. This would raise concerns regarding model generalization and model overfitting. The previous studies have shared one approach which is focusing on studying one aspect of the multifaceted hate language behavior (e.g. racism or aggression). Following the same approach, the hate datasets were designed and crafted to focus on a specific aspect of hate language. However, this focus makes it limited and difficult to identify general hate language across various events on social media. Authors in [171] investigated the abusive language generalization across datasets of different abusive focuses. Their findings and observations concluded that models trained using datasets with a broader coverage of phenomena are more robust in capturing a wider range of abusive language contents. Similar observation was found in the work [228] where authors claimed that supervised learning using domain-specific datasets performs poorly on cross-domain datasets due to the reason that they are attached to the domain-specific information. Their results demonstrated the effectiveness of using domain-independent abusive lexicon to detect abusive language in cross-domain social media datasets. Waseem etal. [225] confirmed the possibility of obtaining high-performance models to detect hate and abusive language when built using composite datasets. However, no considerations have been given to OSNs-specific features such as exclamation marks, words with repetitive characters, or iconic emotions (i.e. emojis and emoticons) that are capable of emphasizing the literal meaning of messages or even reversing it [53].

The abovementioned limitations and current practices of generating specific resources for every domain independently for social media analysis, requires high cost and extensive efforts in term of resources, time, effort, domain experts, and human labor. Aiming to overcome the above limitations and to complement the existing OSB datasets, This thesis

proposed two approaches to address all the mentioned issues. First, it proposes a new domain-independent dataset that follows a high quality protocol for domain-independent data collection and annotation. The domain-free multimedia sentiment dataset (DFMSD) [1] was constructed for multimedia sentiment behavior analysis. It is to note that the choice of sentiment behavior to construct our dataset is due to the fact that the sentiment represents a seed for many further applications and research domains such as user behavior analysis and prediction [125]. The DFMSD dataset considers the subjective and objective aspects of sentiment (i.e. positive, negative, neutral). Also, it provides a decent data size and a high quality psychologist-based manual annotation and the DFMSD was constructed free of restrictions to any domains or keywords as an attempt to bridge specific domains mismatches [19]. It also can be used to enhance learning of sentiment in the domain-specific problems by transferring the general sentiment knowledge instead of starting from scratch. To the best of our knowledge, this thesis presents one of the first studies to build a general (i.e. domain-independent) sentiment datasets for online social media analysis. Second, this thesis proposes to combine different phenomena of an OSB as an attempt to expand the model learning to a wider and deeper patterns of the corresponding OSB of interest. This approach was applied to the toxic speech behavior analysis on OSN as case study in this thesis. The findings in [225, 228] support our decision to combine different phenomena of toxic speech in order to expand the scope of toxic behavior identification on social media.

2.2.2 OSB Modeling

2.2.2.1 Traditional Machine Learning Approach

The interest in sentiment analysis keeps increasing among researchers as it represents a seed for many further research domains [43, 125] such as fine-grained emotion analysis, psychological human needs analysis, and smart cities. Many sentiment analysis works have been done on both long texts (i.e. document-level) [177] and short-texts (i.e. message-level) [218]. The average length of a document-level text is 241 tokens in IMDB dataset [234]. Unlike a long text, the short text has an average length of ≈ 81 tokens which is the average length that we have observed in our dataset. In this work, we focus on analysing short tweets. We believe that short texts provide concise expression and require lower features space than long texts.

For text classification, the performance of classifiers is highly dependent on selected features. The right features will guarantee good learning output. In text classification, BOW using TF or TF-IDF is the most popular feature. Its effectiveness directly depends on the quality of the dataset it was derived from. Most of the sentiment classification studies use BOW as one of the features to build their models [161, 177, 196]. Since BOW ignores the order of words which in turn ignores the context of texts, n-gram techniques provide a partial solution to the lack-of-context problem [129, 212]. It has been shown that using BOW and n-grams features is insufficient for sentiment learning [8, 214]. Considering specific features containing or representing opinion information has proven to better improve the sentiment learning than when only BOW is used. Authors in the study [124] showed that linguistic feature has enhanced the learning performance over BOW feature

on MVSA dataset. Similarly, frequency of POS feature has demonstrated a better classification performance than BOW feature when trained on Latent Dirichlet Allocation (LDA) algorithm [214].

Although individual features such as sentiment lexicon or BOW are necessary for sentiment learning, they are far from enough to yield good results [128]. Integrating BOW with sentiment-rich clue features has shown to be more effective in the sentiment analysis [25]. The integration of emoticons with BOW as proposed in the study [219] has boosted the sentiment learning performance by 13% than when using BOW feature only. Aloufi and Elsaddik [8] showed that combining sentiment lexicons and POS features with the BOW feature also improves the sentiment learning process. The use of sentiment lexicons is shown to be necessarily informative for the sentiment classification [148] especially for the minor classes in cases where the classes are imbalanced. The results provided by Zhu et al. [161] showed that using SentiStrenght sentiment lexicon yielded better performance than BOW-TF for the minor class. When using sentiment lexicons for training sentiment models, we actually combine two learning approaches as suggested in literatures [7, 56, 114, 124]: (1) statistical machine learning approach and (2) lexicon-based approach. When using features other than BOW, the occurrence of feature and frequency of occurrences [81, 148] are the two popular approaches to use in sentiment analysis. We adopt the two approaches in our proposed features.

Previous works on sentiment analysis have focused mostly on support vector machine (SVM), naïve -bayes (NB), logistic regression (LR), random forests (RF), and decision trees (DS) to build sentiment classifiers [125]. The reason that they are the most applied classifiers is due to the better performance they provide in comparison to other classifiers such as k-nearest neighbour (KNN). Pang and Lee [161] used NB, maximum entropy (ME), and SVM to learn sentiment from texts. The result showed that SVM was the winner among the other classifiers. Bilal et al [35] conducted a similar sentiment analysis research using NB, DT, and KNN algorithms. NB classifier has shown better performance than DT and KNN methods. Another sentiment analysis work done by Wan and Gao [218] on Airline Service twitter dataset, showed that RF outperforms NB, SVM, Bayesian Network, and DT when conducting binary classification while DT outperforms the others when training on three classes. Also, four classifiers were used in training a binary sentiment model in the work [212] and the results showed that SVM was the winner among NB, ME, and stochastic gradient descent (SGD). Recently, deep learning algorithms have achieved very good results in sentiment analysis domain [231]. It differs from machine learning in its ability to learn features directly from data. However, explainability of features and learning can be heuristically understood [61]. In contrary, machine learning along with feature engineering, would easily offer such explainablity and interpretability of learning and feature importance especially for unstructured texts. The availability of sentiment resources created by domain experts, makes it easier and faster to craft features and hence reduce the computational complexity of deep learning. In this work, we propose to use machine learning along with feature engineering in order to understand our dataset (DFSMD) and to provide explainable evaluation of its quality on the sentiment learning.

Recently, researchers have proposed to use ensemble classifiers (a combination of multiple classifiers) to build more accurate sentiment classifiers for textual contents on OSNs

[116]. It has been found that ensemble methods are effectively capable of scaling out as data volume increases. In Lin and Kolcz research [127], individual models are trained independently and then evidences from each model are combined for the final prediction. Predictions from the ensemble method have been shown to be better than predictions from individual classifiers. This type of ensemble uses bagging approach and is based on taking the majority votes of all the classifiers [60]. Another type of ensemble uses boosting approach that utilizes the weighted average to build a strong learner from weak ones [190]. Adaptive Boosting (AdaBoost) and Gradient Boosting (GBM) are the most common techniques of the boosting ensemble. In this thesis, we have used the GBM method, as it proves to well handle the high dimensionality and high sparseness problems [26], which is an advantage in the case of sentiment analysis. We have experimented with two powerful GBM algorithms: (1) Extreme Gradient Boosting (XGB) [50] and Light Gradient Boosting Machine (LGBM) [111]. Authors in [105] proposed to use XGB with lexical and embedding features for emotional analysis of tweets. Combining the XGB model with the convolutional neural network model has shown an improvement in the overall performance of the proposed system. The capabilities of XGB to cope with large-scale data has allowed the ensemble model to improve its overall performance. Another work [110] proposed to build an XGB sentiment classifier for financial news and headlines. When training on combination of uni-gram and bi-gram feature, the XGB model has shown to be more effective than when training on other features of TF-IDF and paragraph vector features. Again, a sentiment model learnt using XGB algorithm has shown to outperform other algorithms (i.e. SVM and Gradient Boosting Trees (GBT)) when evaluating Telugu news collected from news websites [149]. The model was trained to recognize the polarity (i.e. positive and negative) of news texts in Telugu language. An LGBM model was trained on telephone conversations data for the purpose of finding the sentiment intensity of the conversations. The LGBM model showed a powerful advantage with 4% better performance than LR model on a combination of text and audio data. TF-IDF was the only textual feature used to train the LGBM. Fan et al. [69] built a sentiment model for recognizing the opinions (i.e. positive, neutral, and negative) of English national team fans during FIFA World Cup 2018. They trained LR, XGB, and LGBM models independently on tweets using word-based and character-based TF-IDF features. Then they calculated the weighted average of all the three predictions from the three proposed models as the final predicted result. The results were promising and showed that the sentiment peaked when the England team were scoring victorious. However, the work did not report the performance of individual classifiers; instead, it reported the result after combing all the three classifiers in terms of weighted performance average. In our work, we propose to train both XGB and LGBT using five types of features. To the best of our knowledge, this thesis presents one of the first works to build a general sentiment (i.e. domain free) model based on LGBM using our domain-free dataset (DFSMD).

2.2.2.2 Deep Learning Approach

Traditional machine learning requires feature engineering as an essential prestep to design, extract, and select a set of chosen features. The quality of model learning directly depends

on the quality of the selected features. Therefore, domain knowledge is essential in deciding features that have importance to the problem in hand. This actually adds a dimension of challenges ranges from finding domain experts to the complexity of translating knowledge into data features. This complexity is specially true for the uncontrolled nature of data coming from an open-platform online social media. Open platform encompasses various domains, topic, and different languages and even dialects. This very openness leads to resource insufficiency, which in turn makes it significantly harder to craft features for every task in every language and dialect.

The advancement of deep neural networks has led to a substantial improvement in Natural Language Processing (NLP) tasks including sequence classifications [87] and computer vision tasks [18, 122, 123, 221, 239]. Neural networks come with the capacity of mitigating the complexity of feature engineering and provide self-learning of data or feature representations. While memory neural networks with attention mechanism have been widely used to capture the sequential information of texts [49], CNNs [24, 100] have been popular for solving image recognition and classification problems [24,123]. Comparing transfer learning in computer vision to that in NLP few years ago, transfer learning in computer vision was by far more successful in performing computer vision tasks. Earlier NLP efforts had been put to exploit previous knowledge by using textual embeddings [38, 174] to avoid restarting training from scratch. Although these embeddings were trained on huge volumes of data, they still suffer from context-independence problem which means that word representations are the same regardless of their surrounding context. More recently, transformer-based language models such as BERT [112] and GBT-2 [176] have made a groundbreaking milestone in transfer learning in NLP. These models are capable of alleviating the complexity of feature engineering and overcoming the limitation of context-independence issue. BERT has achieved state-of-the-art results in learning semantics of textual expressions for various problems including sentiment [63, 204, 222] and hate speech analyses [171, 225, 228].

Recent studies on using BERT models for sentiment analysis have focused on domain-specific analysis [63, 152, 178, 179] and the polarity aspect of the sentiment while ignoring the objectivity part of texts [63, 152, 204]. Moreover, training and evaluation have been conducted on small datasets [63, 204] and many works have not handled the OSNs cultural language such as iconic emotions (i.e. emojis and emoticons) [27, 79, 152]. It has been repeatedly reported that training deep neural networks using large datasets yields better results than the training using small datasets [64]. Also, iconic emotions contain sentimental clues that would greatly contribute in sentiment learning [19, 99]. In addition, automatically or noisy annotated data has been used to train BERT-based sentiment model [27, 104] which in turn compromises the knowledge quality of the learning process. Even though cross-domain sentiment learning has been addressed, many studies focus on adapting sub-domain to one another while considering the main domain to be the same [63, 153, 178]. This approach would not generalize well on various domains since those models will latch on to domain-specific information [228]. This thesis addresses all the mentioned issues. First, it considers the subjective and objective aspects of sentiment (i.e. positive, negative, neutral). Second, it provides a training dataset (DFSMD) of a decent size and of high quality psychologist-based manual annotation. Third, it provides a domain-free dataset (DFSMD) that was constructed free of restrictions to any domains or keywords. Fourth,

it proposes a domain-free BERT-based sentiment model to bridge specific domains mismatches [19]. It also can be used to enhance learning of sentiment in the domain-specific problems by transferring the general sentiment knowledge instead of starting from scratch. To the best of our knowledge, this thesis presents one of the first studies to build a general sentiment model based on BERT language model.

BERT models have shown state-of-the-art performance in detecting hate speech on social media [53, 146]. Marzieh et al. [147] trained BERT-based models for different hate speech categories: racism/sexism and hate/offensive. The overall results showed that BERT-based models yielded excellent performance. Offensive language was studied in [160] using BERT pre-trained model as a base for modeling offensive classifier using OffensEval-2019 dataset. BERT-based model was shown to outperform classical machine learning methods in identifying offensive language in tweets. However, iconic emotions (i.e. emojis and punctuation-based emoticons) were ignored. OSNs specific feature such as exclamations marks, words with repetitive characters (e.g looool) were ignored even though they contain strong sentimental insights [19]. Authors in [189] targeted the problem of aggression and misogynistic identification for three languages on social media. Their approach included using BERT pre-trained model yet no fine-tuning was conducted. They reported that BERT-based model had shown a better performance on binary classification than that of multi-class classification. Hind et al. [4] proposed using domain-specific word embedding with BERT model to detect white supremacist hate speech if it existed on social media. The evaluation was conducted on balanced dataset collected from Twitter and Stormfront forum. However, the size of the proposed training set is quiet small (i.e. 4588 messages) to be able to fine tune BERT architectures. This would raise concerns regarding model generalization and network's overfitting.

The previous studies have shared one approach which is focusing on studying one aspect of the multifaceted hate language behavior (e.g. racism or aggression). Following the same approach, the hate datasets were designed and crafted to focus on a specific aspect of hate language. However, this focus makes it limited and difficult to identify general hate language across various events on social media. Authors in [171] investigated the abusive language generalization across datasets of different abusive focuses. Their findings and observations concluded that models trained using datasets with a broader coverage of phenomena are more robust in capturing a wider range of abusive language contents. LSTM-based models were shown to outperform models based on linear support vector classifier. In this work, we propose exploiting transfer learning using pre-trained BERT model as well as OSN-specific emotion hints like iconic emotions, in order to build our hate classifier.

2.2.3 OSN-based Multilingial Multidialectal OSB Understanding

2.2.3.1 Language Machine Translation

Th advent of social media has revolutionized the language that people use for communication; the informal nature of conversations and communication has become the new norm

on online social media platforms like Twitter. This is plainly evident in the Arabic language where the urban dialect has become the dominant communication language instead of Modern Standard Arabic (MSA) that is different from the dialectal Arabic in terms of morphology, lexicons, and expressions. As a result, existing translation systems designed for MSA would fail to work well with Arabic dialects. In light of this, it is necessary to adapt to the informal nature of communication on OSNs by developing translation systems that can effectively handle the various dialectal Arabic language, besides the MSA. Unlike MSA that shows advanced progress in translation systems such as Google Translate, little efforts have been exerted to utilize Arabic dialects for machine translation systems. The main limitation contributing to the immaturity of Arabic dialect translation systems is the insufficiency of data resources and datasets. Only few efforts have been made to create contents and build datasets for dialectal Arabic. The Bible was among the first data resources that was translated into Arabic Moroccan [3] and Tunisian dialects [4]. In the context of online data, Zbib et al. [237] created Egyptian-to- English and Levantine-to-English dataset collected from dialectal Arabic weblogs and online user groups [5]. The translation from Arabic to English was carried out through crowdsourcing technique on Mechanical Turk by Arabic users. Bouamor et al. [41] used a subset of 2K sentences from Zbib's dataset and extended the translation to Palestinian, Syrian, and Tunisian dialects. The authors asked native dialectal Arabic speakers to do the translation from Egyptian sentences to their own native dialects. Later, Bouamor et al. [42] created parallel-phrase and parallel-sentence datasets that cover various city-level Arabic dialects. The corpus was created by translating a subset of phrases and sentences taken from the Basic Traveling Expression Corpus (BTEC) [208]. Even though those Arabic dialect datasets were created as an attempt to support the Arabic multi-dialect translation systems, they reveal a number of issues that might limit the translation performance of Arabic multi dialects systems on online social networks (OSNs). This claim is based on nine observed reasons: (1) the size of translated sentences to each dialect is small; some datasets are as small as 2K per dialect [42], (2) translation was done by non-professional translators and it is unknown if the translators were native in Arabic dialect and native (or at least fluent in English) to ensure accurate translation [237], (3) translators were not checked if they were familiar with informal and slang English and with OSNs language (i.e. active on social media), (4) some datasets are domain dependent [6], [208] and might not be adapted to other domains, (5) idiomatic expressions were not taken into consideration, (6) most of the datasets did not consider identical sentence translation to different dialects, (7) ONSs cultural expressions were not included in the datasets nor translation, (8) code-borrowing terms were not taken into consideration, (9) the translation criteria and guidelines for most of the datasets have some ambiguity or even absent [41,42,237]. Aiming to overcome the above limitations and to complement the existing Arabic dialectal datasets for machine translation on social media, we propose an OSN-based multi-lingual-dialect Arabic dataset (OSN-MLMD). OSN-MLMD is created by contextually translating English tweets into Spanish, French, and four main urban dialects: Gulf, Yemeni, Iraqi, and Levantine/Shami. To the best of

[3]https://www.biblesociety.ma
[4]https://www.bible.com
[5]https://catalog.ldc.upenn.edu/LDC2012T09
[6]https://www.biblesociety.ma, https://www.bible.com

our knowledge, we are the first to construct a multi-lingual-dialectal parallel translation dataset between English, Spanish, French and dialectal Arabic for machine translation that is optimized for social media informal language. Moreover, we are the first to create a guideline framework that is applicable for translating not only a foreign language into different Arabic dialects , but also from a language to any other language and dialect.

Machine translation models trained on MSA data are not capable of performing well in translating from and into Arabic dialects [185]. Hence, research efforts have been made to enrich such models to explicitly handle the translation of Arabic dialects. The very first efforts towards building an Arabic dialectal machine translation were based on statistical learning approach [186, 187]. Some studies adopted the approach of appending dialect to MSA as a preprocessing step [187,187,188]. Other studies adapted MSA-to-English systems to dialectal data [186]. Recently, deep learning methods have become more dominant than statistical learning approach, and neural machine translation (NMT)has shown to outperform STM with the state-of-art results [183]. NMT for Arabic dialects has not been extensively explored; however, some work on translating Arabic dialects using NMT has been recently introduced. Baniata et al. [31] designed a recurrent neural network-based encoder-decoder neural machine translation model that uses multi-task learning with individual encoders for both MSA and dialects and with a shared decoder. Transformer neural networks have become an alternative to sequence-to-sequence neural networks for neural machine translation. Shapiro and Duh [195] trained transformer-based models for MSA, Egyptian, and Levantine dialects. Their results showed that training a model on multidialectal data is able to benefit the translation of unknown dialect. Sajjad et al. [185] found that transformer based NMT models performed better translations when trained on large-scale datasets compared to small sized datasets in the settings of training from scratch. On the other hand, their results revealed that using transfer learning through fine-tuning pretrained NMT models (i.e. on MSA) improved the learning of Arabic dialectal translations. In our work, we use transformer networks and transfer learning in order to train our NMT models using our proposed OSN-MLMD dataset.

2.2.3.2 Visual Multi-Modality OSB Analysis

Visual OSB analysis for social media intends to extract the information related to the online social behavior of interest from the visual content shared by social media users. Visual sentiment analysis is an example of the online social behaviors that has been given attention in research recently. Initial efforts have considered establishing a direct mapping between sentiment orientation and visual features for visual sentiment analysis on OSNs [23, 200]. However, it has been found that the methods that are based on low-level features do not apply to visual sentiment analysis on OSNs due to that fact that images are freely shared (i.e. conveying a wide range of topics from different domains) on social media and that the relationship between the visual data and its sentiment orientation is extremely complex [125]. It is important to mention that the emotional semantics of visual contents shared on social media are indirectly driven by cognitive semantics, hence the usage of low-level features introduces a problem of a semantic gap [125]. Recently, researchers have utilized deep learning methods for visual sentiment analysis [18, 48, 109, 121, 233] to fill

the semantic gap between the low-level features and sentiment orientation. Deep learning-based methods make the sentiment predictions more interpretable as it transforms visual low-level features into an abstract feature space in which they benefit the analysis of emotional semantic for visual content on online social media. Although deep learning-based approach has achieved some progress in visual sentiment analysis on social media, current studies have not sufficiently given enough attention to the objects in visual content [125]. This means that the process of visual perception is ignored while establishing the mapping between image pixels and sentiment orientation. This thesis addresses this gap and proposes to consider two types of objects in images for visual perception: (1) faces to extract facial emotions and (2) texts to extract sentimental hints. Further, the unreliability of sentiment annotations in existing datasets affects the quality of visual sentiment models and hence increases the difficulty of network training. In this thesis, we propose a domain-free multimedia sentiment dataset (DFMSD) [1] that follows a high quality and strict protocol for data collection and annotation.

Facial Expression Recognition (FER) in the wild of online social media (OSNs) is extremely challenging due to the uncontrolled condition of images being shared. In addition to real faces, there exist animated faces that can be seen in images- like memes for example. Also, due to the uncontrolled condition, those faces might come with variant head poses, occlusions, and face deformation and blur under unconstrained conditions. However, in the past few decades, a great progress has been made in FER where different learning methods that achieved good performance are used. Some of the methods like SVM, Bayesian Network, and Neural Networks require a pre-step to extract facial features before they are used for facial emotion classification which adds substantial effort and computational overhead [45, 54, 73, 194]. Deep learning based methods combine both facial feature extraction and facial emotion classification into one single stage and break the dependency on the hand-crafted features [93, 101]. Convolutional Neural Networks (CNNs) have a natural inductive bias for learning feature representations from images, and thus have shown a promising performance in FER [122, 123, 221, 239]. However, the CNN-based models can be sensitive to image occlusion complex backgrounds, or variant head poses to name few [120]. Recent studies have shown that vision transformers(ViT) are robust against image occlusion and disturbance [102, 132, 155] which justifies our decision to use ViT as the backbone of our FER model on OSNs.

The novel transformers [62] have become the state-of-the-art method in NLP tasks, and recently it has been applied in computer vision tasks [95, 134]. Visual Transformers have achieved a remarkable performance in image classification tasks [62, 139] and outperformed CNNs in terms of computational efficiency and accuracy [62]. ViT was designed based on the attention mechanism which has proven to be a key element for image classification to achieve high performance robustness. ViT uses the attention mechanisms directly on a sequence of input image patches without depending on CNNs where attention is either used in conjunction with it (i.e. with CNNs) or to replace some components of the CNNs. When trained on enough data, ViT outperformed the performance of similar state-of-art CNNs with four times fewer computation resources and four times better efficiency and accuracy [62]. Unlike CNNs, which have small local receptive fields in each layer, the multi-head self attention layer allows the ViT to embed information globally (i.e., attend

to global features) across the overall image. Moreover, the model learns to encode the relative location of image patches so that it reconstructs the image structure.

ViT has shown to perform the best when trained on large-scale data; that was manifested in its performance on ResNet against ImageNet [62]. This means ViT is able to generalize well on image classification tasks when trained on large-scale data compared to when trained on small datasets. Researchers [18,120] benefit from such vision transformers models trained on large-scale data by exploiting transfer learning approach that uses pretrained weights to finetune transformer architecture on smaller datasets. Authors in [120] have adopted ViT transfer learning to learn an FER model through one stage finetuning with using pretrained weights from a transformer based Deit-S model. Ma et al. [131] have shown the positive impact of using pretrained weights (i.e. obtained from ImageNet-21K) when training their transformers based FER model compared to when the training was conducted from the scratch. While there have been vision transformers based FER models, to the best of our knowledge we could not find studies that apply vision transformers on visual sentiment classification tasks. Transfer learning has addressed the problem of small datasets (i.e. given the fact that it is time, cost, resource consuming to build large-scale manual annotated datasets) for image classification problems including FER and sentiments. To further compensate for the small datasets, existing studies suggest overcoming the difference between source task and target task when using transfer learning, through a two-stage finetuning strategy [113, 156]. First stage finetuning shifts the learning from the source task to the target task while second stage finetuning refines the learning in the target task [140]. The two-stage strategy has shown to outperform the one-stage finetuning on FER tasks that use small datasets [140, 220]. In this work, we adopt the two-stage strategy to build our ViT based FER and sentiment models.

Several efforts have been made to analyze the sentiment and emotion using textual and visual modalities [168]. In the context of textual sentiment analysis, Alzamzami and El Saddik [17] attempted domain independent sentiment analysis using DL transformers network and showed that their model is able to adapt to various domains including sport and movie reviews. Image sentiment recognition is an area of research interest as well. Sun et al. [205] designed an algorithm that discovers affective regions and supplementing local features in images top boost the performance of visual sentiment analysis. Multimodal emotion and sentiment recognition has had an equally active research area in the last few years. P. Fortin et al. [169] proposed a multimodal architecture for emotion recognition system that performs predictions in the absence of one or two modalities by using a classifier for each combination of text, image, and tags. Another work by Nan Xu et al. [201] where an interplay of visual and textual content for sentiment recognition was modeled based on a co-memory network.

The learning of multimodal sentiment recognition requires a feature extraction method and fusion strategy [65]. Most of the previous work on multimodal emotion and sentiment recognition use low level features (e.g. SIFT for visual modalities and Glove for textual modalities) or deep features [65]. Features that are extracted a pre-trained deep learning models are called deep features. They are extracted after a DL model is trained using a labeled dataset. Existing studies on facial recognition have extracted deep features from pre-trained facial recognition networks, and similarly pre-trained text deep models have

been used to extract text features for emotion and sentiment and analysis [65]. Such studies highlight that the deep features yielded better performance compared to low-level features. CNNs pre-trained models are widely used for deep feature extraction due to its natural inductive bias. Authors in [169] used DenseNet-121 pretrained on ImageNet to extract deep features for image, text, and tag models before they concatenate these features and feed them to a two fully-connected layers for final predictions. Similarly, authors in [115] used CNNs based pretrained models for deep feature extraction; VGG16 for image features and BalanceNet for textual features. A hybrid of intermediate and late fusion approaches was implemented based on CNNs to concatenate the features for the final sentiment predictions. In this work, we follow multimodality approach in [115, 169] and propose to use three transformer based pre-trained deep models to extract features for image, facial emotion, and text. We adopt the intermediate fusion approach to fuse image, facial emotion, and textual features and feed them to an MLP architecture in order to build our multimodal sentiment classifier. To the best of our knowledge, we are the first to use transformer based fusion approach with three pre-trained transformers based models to extract features for multimodal sentiment analysis on OSNs.

2.2.4 OSN-based Data Exploration and Dynamic Interpretation

Bag-of-words (BOW) and Term-Frequency-Inverse-Document-Frequency (TFIDF) features at the n-gram level have been widely used with LDA and NMF algorithms for topic modeling in social media [198, 206] particularly in COVID-19 related social analysis [14, 119, 167]. Many of these studies have only accommodated removing canonical stopwords (e.g. "the", "and") during the pre-possessing step to construct features. However, removing canonical stopwords does not entirely solve the problem of the existence of common uninformative words, which will definitely affect the quality of the topic models. For example, LDA models trained without removing common words will produce topics with high probabilities of uninformative words. To overcome this issue, literature suggests removing domain-specific [193] and corpus-specific stopwords [66]. Such methods have been proven effective in enhancing coherence across topics. Authors in the work [137] took it further and showed that lemmatizing the corpus and limiting the vocabulary of news collections to only nouns, has improved the semantic coherence of topic models. However, reporting news is one part of social media data. OSNs platforms are open; hence the data flow spectrum is broad ranging from reviewing a product, expressing frustration, to reporting news. Ignoring other part-of-speech tags will result in throwing important information that can be found in nouns and verbs for example. Incorporating different part-of-speech tags for topic modeling [57] has shown to produce reasonable topics on a small Twitter dataset. In this thesis, we apply the same approach but on a large scale datasets.

While topic models are proven effective in extracting latent patterns (i.e. themes or topics) out of social media data [107], they fall short in providing human-friendly interpretations for these topics [85]. Manual interpretation of topics is subject to human bias [138]. Moreover, given the diversity of OSNs contents and huge data volumes makes the availability of domain experts to annotate data for various problems, a difficult task. Early researches on topic labelling focused on exploiting external knowledge resources in order

to automatically label topics of topic models. However, this approach is not applicable to OSNs data streams since the emerging social contents and events discussed in OSNs might not exist in these external resources in a timely manner [32]. Later, the focus redirected towards labelling topics with the most representative single words based on the output of topic models [136]. Single words provide generic meaning, which makes it difficult for users to create the main idea when single words of topic models are combined. In addition, single words may often be homonyms (i.e. they sound the same and have the same spelling but do not have related meanings) or polysemous (i.e. the word is used to express different meanings depending on the context). In this context, Qiaozhu et al. [138] proposed the use of phrase labels to automatically label LDA-style topics. The results of their questionnaire showed that people prefer phrases over words for topic comprehension. However, their approach depends on NLP techniques (i.e. chuncking, POS tagging, and n-gram) which is resource and time consuming. Additionally, the approach focuses on topics derived from static well-formatted documents (i.e. news articles and scientific article) which is the case in the work [92] as well. Recently, Amparo et al. [32] have tackled this issue and presented the topic labelling of Tweets as a summarization problem. The results demonstrated that the topic labels generated by their method showed that the use of summaries, as topic labels outperformed the use of top n words resulting from LDA model. However, the output summaries consist of single words computed using methods based on TextRank and TFIDF where the latter was shown to yield the best labels. Given the dynamic size of topics ranging from being small to large, TFIDF would fall short on small data sizes. Recently, a phrase-based topic labelling approach [74] has been developed based on OSNs activities parameters (i.e. views and likes). However, phrases of length two were only considered. The meaning of a sentence varies with the order and length of its constituting words (e.g. noun-verb-adjective phrase). In this work, we propose to use RAKE algorithm for automatic topic interpretation as it solves the mentioned issues of current topic labelling methods. To the best of our knowledge, this work is the first to address these issues and to utilize RAKE algorithm for automatic interpretation of LDA-style topics using OSNs data.

Chapter 3

Multimedia Multi-Lingual-Dialectal Online Social Behavior Framework: An Overview

Figure 3.1 illustrates our proposed framework. The framework is designed and developed to model domain-independent online social behavior through different languages and dialects on social media contents. Its applications could be widely utilized with diverse contexts. Based on the application objective, data of interest will be crawled from a social media platform (i.e. Twitter for this work). The collected data includes textual and visual contents that will go under cleaning process before it is saved and set in an appropriate format for the preprocessing and feature extraction steps later on. The online social behavior will be recognized and interpreted automatically using the developed models that will be explained through an overview description of the main components. Then, the data analysis engine will take the results from the main components and produce an analysis of interest be it temporal based, geo-location-based, language/dialect based, or topic-based analysis.

The design and development of domain-independent online social behavior analysis are discussed in Chapter 4. The details of developing multi-lingual-dialectal online social behavior models are presented in Chapter 5 while the process of designing and implementing a visual online social behavior model is discussed in Chapter 6. Chapter 7 presents the development of multilingual topic modeling and dynamic topic interpretation.

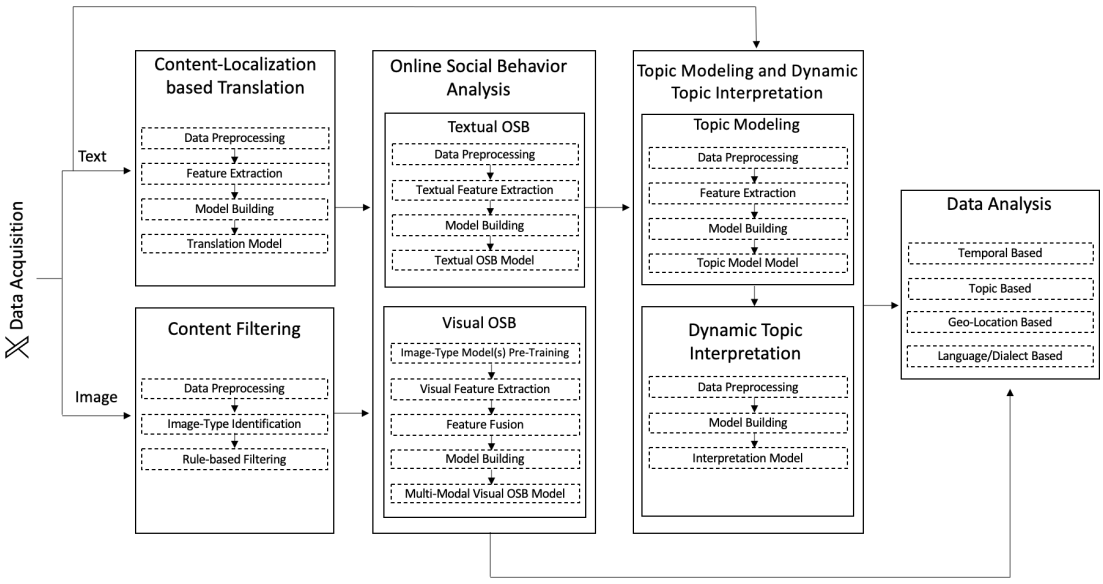


Figure 3.1: **The proposed multimedia multi-lingual-dialectal framework for modeling domain-independent online social behavior.**

3.1 Content-Localization based Machine Translation

Two third of the users on OSNs use languages other than English to access and publish posts and contents. Further, the informal nature of social media conversations appears to be the most common form of communication given that informality differs widely between languages in form of dialects. Arabic language is an example that its dialectal form is used more often than standard form on social media. The complexity with Arabic language is that it has many different dialects varying not only from country to country but also from region to region within the same country. This results in under-resourced status for dialectal languages like Arabic. Given the expensive cost for building resources for each language and dialect and to solve different tasks, we propose using content-localization based translation approach for online conversations of different languages/dialects. Using this approach, we aim at solving the resource insufficiency of an under-resourced language by translating its contents into a high-resourced language and hence be able to use the available resources -in term of data and models- of that language. A comprehensive description of this component is detailed in Chapter 5.

3.2 Multimedia Online Social Behavior Modeling (OSB)

We study -in this thesis- two types of online social behaviors: sentiment and toxicity in OSN conversations (i.e. tweets in this thesis). The decision for choosing those two online social behaviors has come according to the wide array of applications and domains including

cyber security and public health, that will benefit from such the recognition and analysis of such behaviors.

The domain-dependent approach comes with the concern that they do not generalize to other domains [1,17,228] since they latch to the information of the domain they have learned from. Our OSB component, instead, mainly focus on the domain-independent approach to model and analyze online social behaviors in social media. While texts require a knowledge of their spoken language, images speak universal language. Universal language of images refers to visual contents and emotions where no language is needed to interpret its contents. We exploit images for online social behavior analysis as an attempt to provide an ability to extend the analysis of online social behavior especially when there is a shortage and lack of sufficient resources (i.e. datasets and models) in different languages and dialects. The details of developing the proposed intelligent models for online social behavior are described in Chapters 4, 5, and 6.

3.3 Topic Modeling and Dynamic Topic Interpretation

The main objective of this component is to explore and find patterns in social media data and then generate explainable interpenetration of these patterns. Unsupervised learning approach is used in the modeling for this component. This component consists of two main sub-components: topic modeling and topic interpretation. Topic modeling approach is adopted to explore hidden patterns in large datasets and discover latent topic within data without prior human knowledge involved. Topic modeling methods are sensitive to data noises, which is a normal phenomenon in OSNs data. Therefore, the topic modeling adopted for this work is designed with criteria to fit the challenging unstructured and noisy nature of social media data as an attempt to maximize the efficiency of extracting useful information and recognizing hidden insights from social media data. Topic interpretation sub-component is responsible for automatically generating coherent interpretation of the inferred topics resulted from topic models since manual topic interpretations requires human efforts and can be easily biased towards subjective opinions [136]. The interpretations are decided to be generated in a dynamic-length phrase format. The reason for this decision is that single words create difficulties to comprehend the main meaning of topics while sentences are too specific and might miss other aspects of topics. In this work, we use the topic modeling and interpretations as complementary tools to facilitate the understanding of online social behavior analysis. A comprehensive description of this component is detailed in Chapter 7.

3.4 Data Analysis

The data analysis engine uses the outputs of OSB and DEI engines and generates an analytic story from different views: temporal and topic-based, geo-region, and language/dialect analysis. In temporal analysis, the online social behavior is illustrated in a time-line manner (i.e. over days, months, seasons, etc). Theme-based (i.e. topic-based) analysis provides

non-linear analysis that is based on the themes and patterns found throughout the given datasets. Geo-region based analysis illustrates the analysis based on the geo-region that the OSN messages generated from. Finally, language/dialect based analysis provides a cultural view based on the spoken language/dialect of OSNs conversations.

Chapter 4

Domain-Independent Online Social Behavior Modeling

4.1 Introduction

This chapter presents the definition of online social behavior studied in this thesis, followed by the details of design and implementation of the textual sub-component of the domain-independent "Online Social Behavior" component illustrated in Figure 3.1.

Sentiment, emotion, sarcasm, and humor to name few, have been studied individually in existing works. However, - to the best of our knowledge - there is no existing work that categorize them all together under one umbrella and provided a descriptive standard definition for them.

People, in real-life settings, interact using means of symbolic-based cultural systems like language in which the shared rules of language and other symbolic systems govern these interactions or communications [192]. Communication in turn could be defined to "include the information that people acquire, through inference, from all kinds of interactors participating in material medium" [192]. The entire set of communications taking place during an interval of time with reference to a group of people can define what it is called a behavioral system [191]. Therefore, "a behavioral system is composed of all people-artifacts, people-people, people-extern, artifact-artifact, and artifact-extern interactions relating to the members of a specific household or community or society" - Michael Schiffer [192]. An operation of behavioral system could be looked at as a repetition of its constituent interactions [192]. In this context, Schiffer [192] claims that communication and behavior are related to each other, and both consists of people-people, people-artifact interactions. Similarly, Hannema [90] claims that "when we interact with others, anything we do communicates". Behavior is communication and communication is behavior." Further, Schiffer distinguishes between human behavior seen in stimulus and response and behavior's causes such as goal and intention.

People migrating their real-life communications into digital format on social media where we witness individuals interacting with each other and virtual entities within same or

different geo-locations in different times. Projecting these definitions of communication and behavior on the virtual social world where people interact with people and entities in virtual medium, we can observe a repetition of interactions that share common characteristic. According to Schifer, a human behavior is formed of its constituent interactions. Therefore, a repetition of a set of interactions that share common characteristic form a behavior. Since the focus of this work is on the communication (i.e. textual and visual posts) on social media, we refer to the behavior as online social behavior (OSB). Online social behavior (OSB) can be seen in the sentimental and emotional feeling expressed towards, for example, a soccer match or in language politeness of students' conversations on virtual campus to a hate speech voiced against a certain group. There exists a number of online social behaviors that have been studied in the literature, however, most of them follow on the domain-dependent approach. The domain-dependent approach comes with the concern that they do not generalize to other domains since they latch to the information of the domain they have learned from [1, 17, 228]. Our OSB component, instead, mainly focus on the domain-independent approach to model and analyze online social behaviors in social media. We study two types of online social behaviors: sentiment and toxicity in OSN conversations. The decision for choosing those two online social behaviors has come according to the wide array of applications and domains including cyber security and public health, that will benefit from such the recognition and analysis of such behaviors. In this thesis, we present two types of online social behavior: sentiment and hate. Note that this chapter presents the details for textual online social behavior modeling while the visual online social behavior modeling is discussed in Chapter 6.

The trend among researchers is to build online social behavior models for each domain independently; sentiment is one example of this trend [7, 110]. This actually is very costly due to the following reasons: (1) Data collection needs to be customized to the target domains. This actually is very costly due to the following reasons: (1) Data collection needs to be customized to the target domains. This requires extensive efforts to search and match the required data. (2) Data annotation needs domain experts. It is very hard to have experts agree to annotate a large volume of data, and even if they do, the cost in terms of time and expenses is high, let alone the tediousness of the task. (3) Individual OSB models share the base knowledge (for example in sentiment behavior: like, dislike, love, hate) regardless of the domains they fall under. "Ronaldo was a disappointment in today's match" and "Trump is such a disappointment" reflect negative sentiment even though both sentences are from different domains. Therefore, there is redundancy in preprocessing and training sentiment models for different domains. Also, the domain-dependent approach comes with the concern that they do not generalize to other domains [1, 17, 228] since they latch to the information of the domain they have learned from. The challenges of the domain-specific OSB modeling shed the light on the importance of generalizing the OSB learning independently of any domains. This general knowledge of models would, in turn, help in speeding up the learning process of specific domains or the process of domains adaptation. Instead of learning a model from scratch, the prior knowledge learnt through general OSB would act as a base knowledge to start from there. Another advantage for building general OSB classifiers is that reasonable performance in terms of resources and training can be obtained at a lower cost than building a model for every domain. In this

work, we focus on adopting the domain-independent approach in modeling and analyzing our online social behaviors. The nature of data shared on Twitter is different from the nature of reviews which tend to be predominantly negative or positive. In this thesis, we consider two techniques to solve the domain-independent OSB:

1. Constructing a domain-free OSB dataset (i.e., sentiment in this thesis). Our proposed dataset (DFSMD) [1] was collected free of predefined filters and annotated based on the perspective of the owners of social media posts and not the annotators' point of views.

2. Combining various phenomena of an OSB (i.e., hate speech in this thesis). The UN [1] warns that the communications during COVID-19 pandemic could be exploited to instigate discrimination, stereotyping, stigmatization, racism and xenophobia, all of which fall under the umbrella of hate behavior according to several universal definitions of hate speech [2], [52, 164]. Updating its policy guidelines, Twitter has warned against the use of hate language. It disallows promoting "violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease". Accordingly, it is understandable that hate behavior embodies violence, abuse, or harassment language which indicates that hate language consists of multifaceted contexts.

   Recent trends in hatespeech-related researches have focused on specific targets such as racism or aggression [171, 228]. Accordingly, the datasets have also been featured according to the targeted focuses, thus introducing a challenge of identifying universal patterns of hate language across social media. This approach makes hate detection a domain-dependent task. That being said, the requirements to build different models and datasets to capture different hate language phenomena should increase notably with the presence of limited resources of domain experts and high expenses of datasets manual annotation. The challenge of modeling domain-dependent hate language sheds light on the importance of learning general patterns of hate language. This general knowledge of models helps not only in capturing a wide spectrum of hate behavior across social media, but also in controlling and detecting the spread of hate contents regardless of their types. In this thesis, we combined various phenomena of violence and hate languages as an attempt to build a generalized hate model capable of detecting general patterns of hate language on social media.

4.2 Datasets

This section lists and describes the datasets used for modeling, evaluating, and analyzing the domain-independent online social behavior.

[1] https://www.un.org/

[2] https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1135

4.2.1 Domain Free Multimedia Sentiment Dataset (DFMSD)

In this section, we describe the procedure for the construction of the DFSMD, including data collection, preparation and annotation. We propose a sentiment dataset as sentiment is the base for other online social behaviors; classification of mood, sarcasm recognition, and hate speech detection are examples of the tasks covered by sentiment analysis [39].

4.2.1.1 Dataset Collection

Twitter APIs were used to collect our set of tweets. The fetching process was free of keywords, in order to respect our purpose of creating a generalized dataset independent of topics and emotions. The tweets were collected worldwide from five different dates, chosen randomly to ensure that many of the topics and events discussed daily were covered. To ensure that the content of all tweets was appropriate and distinct, we created a filter to exclude tweets with inappropriate, non useful content (i.e. tweets with only hashtags or links), and retweets. The collection process included only English tweets, since English was the common language among participating annotators. Each tweet entity is described with a set of attributes including creation time, message, tweet user, location and an image. Note that not all the tweets necessarily come with images or a location. A total of 70,228 tweets were collected during the data collection.

4.2.1.2 Dataset Preparation

The collected data from the previous section was noisy and contained tweets with non-useful content. As a result, we performed two runs of cleaning. First, we created three cleaners: 1) to remove mentions to ensure text granularity, 2) to remove links since annotators won't open them, 3) to remove duplications. We used regular expressions to ensure the quality of the cleaning process. Second, we manually read and evaluated the tweets and excluded all the tweets with meaningless contents (i.e tweets auto-generated by applications). The same is applied to tweeted images; we excluded the tweets that came with inappropriate or unrelated images. A sample of 39,000 tweets (i.e. out of 70,228 raw collected tweets) was randomly selected for the manual evaluation. Three English speakers participated in evaluating and choosing the useful tweets. Each evaluated 13,000 tweets that had been cleaned using the first cleaning step. A total of 12,800 tweets were selected for the annotation process.

4.2.1.3 Dataset Annotation

Our dataset annotation was designed based on a set of criteria for both annotators and questions. Quality control was considered as a tool to monitor the behaviour of annotators and to ensure the quality of the data being annotated. To access an acceptable number of sentimental opinions on our tweet set, we used a survey method as a way to acquire as many annotators as possible. To ensure the quality of the resulting annotations, we carefully decided on three main aspects: annotators, questions, and tweet subsets.

4.2.1.3.1 Annotators

In order to create a dataset with high quality annotation, we required that all annotators speak fluent English and possess a high level of education. Many studies reflect the fact that the way people express their opinions and perceive other people's opinions are subject to cultural and gender differences [78, 229]. Accordingly, annotators of both genders and of different cultural backgrounds should be considered in order to limit the bias in the annotation.

An email invitation was sent to email list provided by university of Ottawa. The invitation included a questionnaire asking for personal and demographic information. According to our criteria, 58 out of 65 participants were qualified to participate. 24% of the participants are native and 76% fluent English speakers. The participants are of both genders (57% male, 43% female) and of different cultural backgrounds (East Asian 33%, South Asian 38%, Middle Eastern 21%, African 7%, and North American 1%).

4.2.1.3.2 Questions

Each tweet message was associated with a sentiment question on a 3-level scale (positive, neutral, negative). On each tweet message evaluation, there was a confidence question on a 5-level scale, ranging from 1 (i.e. not confident) to 5 (i.e. very confident). If a tweet came with an image or images, the same questions are asked to evaluate the sentiment of the image(s). In addition, a third question is asked to evaluate if the image(s) is related to the tweet, followed by a confidence question for evaluating the answer. The reason for adding the confidence question in our survey is to show the strength of the sentiment evaluation provided by the annotators. This will be used for two purposes: 1) as a factor to resolve disagreements between annotators, if they exist, 2) as an indicator to exclude tweets with unclear sentiment.

4.2.1.3.3 Tweets Subsets

The tweet dataset was divided into 20 subsets. Every annotator was assigned to only one group. Therefore, each tweet group had at least four annotators. Previous studies suggested that average of 4 non-expert annotators are required for annotation tasks [202]. Our strategy of dividing the tweets into groups was to ensure that the annotation process was not tedious for our participants. Therefore, they were able to provide a high level of concentration in order to carefully evaluate tweets and provide consistent sentiment understanding.

In an effort to increase the speed of the annotation process while also ensuring consistency, annotators were provided with a web-based annotation tool. Guidelines were provided to the annotators explaining the annotation task as well as some examples of annotated tweets for reference. The most important guideline focused on asking the annotators to evaluate the feeling/opinion of the authors when they posted these tweets. Annotators needed to register an account on the survey website. We proposed this service

so that the annotators could stop the annotation at any time and resume safely where they had stopped, at a later time. Only one tweet was displayed at a time. The message and image(s), if any, were shown to the annotators along with the questions and the answer fields. A hidden timer started once a tweet was displayed and stopped when the next tweet was shown. We provided a counter for each annotator to give them a sense of their progression. Once annotators were done with their tweets, the results were saved under their identity.

The annotation process was split into three phases where the first and second phases are combined and then followed by the third one. The first phase acts as a qualification phase, intended to test the annotators' efficiency and exclude the annotators whose work did not meet the required standards. An annotator is qualified if she labels in agreement with a domain-expert. All annotators were provided with tt test tweets (i.e. 200 in our case), obtained from a dataset annotated by three expert psychologists [6]. They were asked to annotate the test tweets with one of the three sentiment states: positive, neutral, negative. Their results were then compared with the sentiment labels of the psychologists to measure the qualification agreement $agr(a)$ as shown in equation 4.1

$$agr(a) = \frac{n}{tt} * 100 \tag{4.1}$$

where a is an annotator, n is the number of matched sentiment labels between annotator a and psychologists, and tt is the size of the test tweets.

Table 4.1: **DFSMD statistics**

Labels	**Tweets Number**	**Images Number**
Positive	6,683	4,851
Negative	2,275	966
Neutral	2,983	4,427

In the second phase, the actual annotation was performed in two rounds, giving the annotators enough time to rest between the rounds. In each round, annotators were divided into ten groups, each with at least four individuals. For each group, the assigned annotators were given 640 tweets and asked to answer the sentiment questions (4.2.1.3.2). Note that each group had a distinct set of tweets. To preserve the quality of our dataset, we combined the test tweets with the to-be-annotated tweets, as a quality control technique. This way we ensured the annotator's mental state was constant when evaluating both types of tweets, and hence eliminating possible biases. As a result, a sentiment evaluation with the same standard for both types of tweets could be obtained. The duration timer is another quality control metric we considered in the annotation process. A hidden timer started when a tweet was displayed and ended when the next tweet was shown. Its role was to measure the time annotators took to evaluate tweets, ensuring that they took enough time to carefully review each tweet before assigning an appropriate label.

Based on the qualification agreement measurement and the duration timer results,

three annotators were excluded. They showed a lower agreement rate (less than 50%) with the psychologists, and in comparison to the majority of the other annotators, who agreed on more than 50% of the tweets. The low agreement rate could have been a result of these annotators selecting the labels randomly, which could be evidenced by examining the timers, or caused by language barriers, which could lead to difficulties in understanding the tweet's slang, idiom or context. The other 55 annotators were considered as qualified and could move to the third annotation phase.

The third phase involved the construction of our final dataset, DFSMD. We constructed the DFSMD based on the following rules: the label was assigned to each tweet within the dataset based on a majority vote among annotators in each of the 20 groups. In 1,848 cases, tweets equally achieved sentiment disagreements. For instance, two annotators thought that a tweet was positive and another two thought that it was neutral. In such cases, we resolved this conflict by considering the confidence answers that the annotators provided. If the confidence score of all the annotators of one opinion x_1 was greater than the other opinions x_2 or x_3, then opinion x_1 was the sentiment of the corresponding tweet. A total of 859 (i.e. out of 1,848 tweets with disagreement) were excluded from the dataset due to the inability to resolve the disagreement between the annotators. Table 4.1 shows the sentiment distribution of DFSMD. The DFSMD[3] is released for public use.

4.2.2 Hate Datasets

In order to train our hate speech classifier, we use four available datasets published in previous studies:

- **HatEval 2019** The English tweet dataset [33] was constructed based on women or immigrants as targets of hate speech in this dataset. The tweet annotations did undergo two steps: (1) by non-expert annotators using crowd sourcing mechanism, (2) then two domain-expert annotators reviewed the annotated tweets. The inter-agreement in annotating the dataset scored 83%. The data set contains a total of 13000 tweets, out of which 5470 tweets are labelled hate speech.

- **OffensEval 2019** The dataset [236] was annotated for categorizing offensive/non-offensive language on twitter. The offensive language is defined as insult or threat contexts. If offensive language is directed towards individuals, groups, or others, it is annotated as hate speech. The annotation was done by domain experts using crowd sourcing approach. The dataset consists of 13240 tweets, 4400 of which are labelled offensive and hate speech.

- **Antigoni Dataset 2018** The tweet dataset [76] was manually annotated using the crowd sourcing approach.

  It consists of hate, abusive, spam, and normal labels. The results have shown that there was confusion between abusive and hate labels during the annotation process,

[3] http://www.mcrlab.net/datasets/dfsmd/

so we decided to combine both under the hate label. We removed the spam label which resulted in a total of 60702 tweets; 28587 of which are normal and 32115 are hateful.

- **Waseem and Hovy 2016** The tweet dataset [224] was annotated for racism and sexism types of hate language. The tweets were reviewed by the authors [224] and then by domain experts.

The labels of these datasets were binarized into two labels: hate and non-hate. This approach was adopted in previous studies [4, 171] and it has been proven effective. The number of hate tweets in our dataset is 39593 where the normal tweets are 46753, which brings the overall total size to 86346 tweets.

4.3 Methodology

4.3.1 Traditional Machine Learning Approach

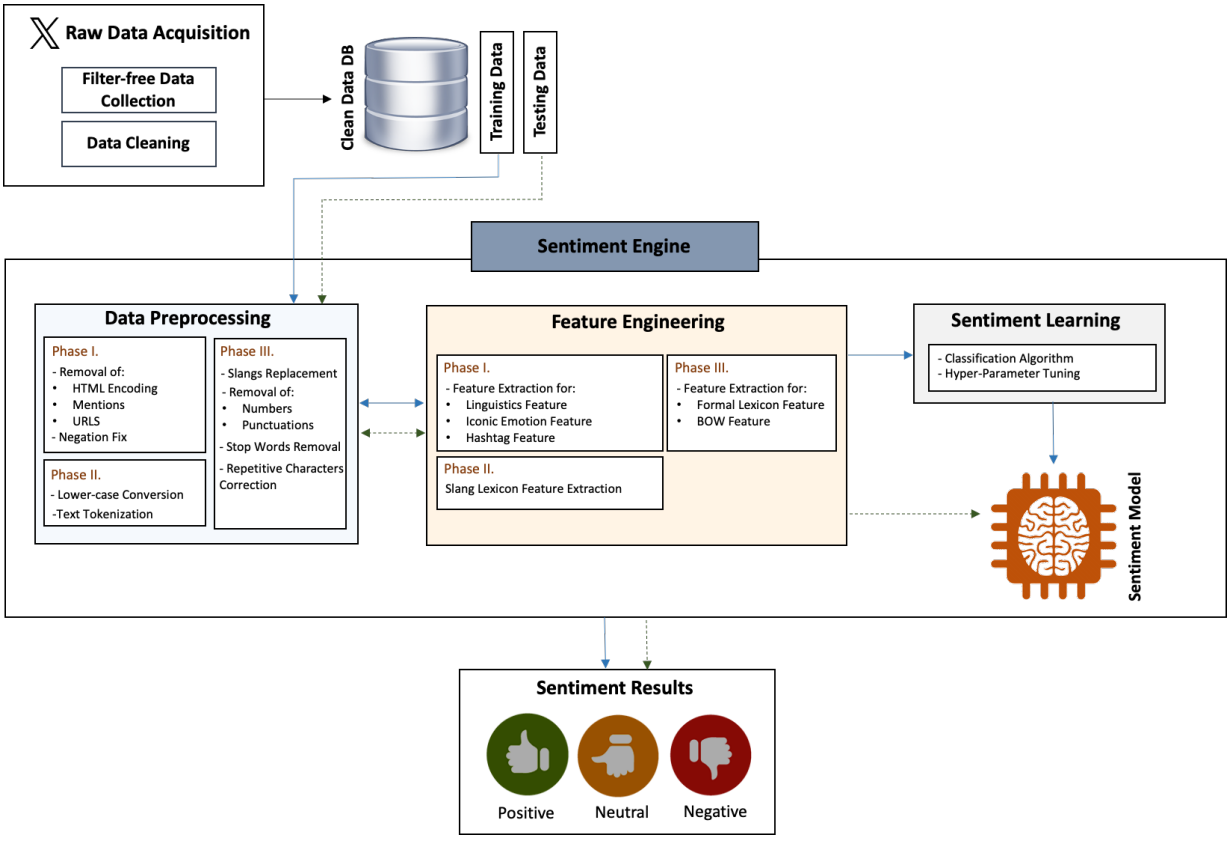


Figure 4.1: **Framework for traditional ML-based online social behavior modeling.**

Figure 4.1. illustrates our framework for the domain-free sentiment classification for short texts. We follow the same framework used in general classification problems. The

data acquisition component is responsible for collecting data based on filter-free criteria. We used Twitter Stream API for this job. Then, the retrieved data underwent a cleaning process to finally generate the Domain-Free Sentiment Multimedia Dataset (DFSMD). Further details on the data collection can be found in our earlier work [1]. Before the data is passed to the sentiment engine, it is split into training and test sets in order to facilitate the learning and evaluation processes. The sentiment engine consists of four components: data preprocessing, feature engineering, sentiment learning, and sentiment model. In the preprocessing phase, the data is prepared based on criteria to keep the important parts of the data that utilize the sentiment learning (explained in details in Section 4.4.). The preprocessed data is then used to extract meaningful features in the feature engineering component, and represented them in a vector format. Both data preprocessing and feature engineering process the data in three phases in order. This is because extracting some features is dependent on the existence of pieces of information that will be removed eventually before the data is fed into the sentiment learning component. A classification algorithm is selected in the sentiment learning component and its parameters are tuned as a prior step to the learning process. While only the training data is fed into the sentiment learning component, both training and test sets are evaluated by the learnt sentiment model. Finally, the result is expressed as a one class of positive, neutral, or negative. The details of the components are presented in the following sections.

4.3.1.1 Feature Engineering

As seen in Section 4.4, sentiment classification on large textual datasets requires a lot of preparation work on the back end. This step is important in order to transform a text into a format that an algorithm can use. The transformation process, which involves representing textual data numerically, is called "feature extraction".

Words and other attributes of text represent either discrete (i.e. frequency of words) or categorical (i.e. presence of words) features. In feature engineering, we aim at mapping these words and attributes into real-valued vectors. We have used different techniques to choose the numerical representations of the textual features.

In this work, we have used different types of features, including Bag-Of- Words (BOW), n-gram, sentiment lexicons, and linguistic hints. We have also used features representing OSNs culture such as iconic emotion, and hashtag. Detailed description of the features is presented in the following subsections.

4.3.1.1.1 Bag-of-Words (BOW)

BOW is a well-known technique in text processing. It generates a list of words, called vocabulary, from a dataset. Each tweet is represented as vector with each word represented with a numerical value depending on the used numerical representation method. For example, a word is given 1 value if it is present in the vocabulary or 0 if otherwise. Another technique is the frequency of occurrences of the words of a text in the vocabulary. The two most common approaches to numerically represent a text are: (1) Term Frequency (TF)

which represents the number of time a word occurs in a tweet with respect to its total number of occurrences in the whole dataset, (2)Frequency-Inverse Document Frequency (TF-IDF) which represents the level of importance of a word in the whole dataset. In this work, we have adapted the TF approach since it shows better performance over TF-IDF during our initial experiments.

4.3.1.1.2 n-gram

BOW ignores the word order, which results in ignoring the context of texts. To solve this, n-gram technique is incorporated to extend the BOW model where a document is represented as n consecutive words [129]. The literature suggests the order of n <=3 consecutive words. In our work, we investigate the impact of using uni-gram, bi-gram, and tri-gram. We use TF approach for our features vector representation.

The n-gram feature vector consists of each n consecutive words in a tweet, as seen below:

- **Uni-gram feature**: the vector will be of m dimension where m is the size of our constructed vocabulary. Each item in the uni-gram vector represents a word from the vocabulary list.
- **Bi-gram feature**: the vector will be of $m-1$ dimension where m is the size of our constructed vocabulary. Each item in the bi-gram vector represents a two-consecutive words from the constructed vocabulary.
- **Tri-gram feature**: the vector will be of $m-2$ dimension where m is the size of our constructed vocabulary. Each item in the bi-gram vector represents a three-consecutive words from the constructed vocabulary.

4.3.1.1.3 Formal-word Sentiment Lexicons

Various resources of sentiment lexicons have been developed so that sentiment learning benefit from textual sources [80]. Each lexicon was built based on a philosophy including but not limited to coarse grained or fine-grained sentiment classification: what part of text to annotate, single-word level or n-gram level. As a result, we propose to use two different lexicon resources: (1) AFINN-111 Lexicon (AFINN) [159]: it contains 2,477 words which were built based on Affective Norms for English Words (ANEW). Each word is annotated with score ranging from 1 to 5 or from -5 to -1 for positive and negative words, respectively, (2) NRC Hashtag Sentiment Lexicon (NRC) [143]: It contains 54,129 words extracted from 775,000 tweets. The tweets are automatically labelled based on the polarity of hashtag such as "amazing", and "terrible".

We have extracted two features, based on the presence of words from our tweets, in the used lexicons. Each feature represents the frequency of positive and negative words, respectively, for each individual lexicon.

For AFINN lexicon, the features extracted are:

- **Affin-positive-feature**: contains the frequency of positively scored words (i.e. in a tweet) which exist in Affin .
- **Affin-negative-feature**: contains the frequency of negatively scored words (i.e. in a tweet) which exist in Affin .

For NRC lexicon, the features extracted are:

- **NRC-positive-feature**: contains the frequency of positively scored words (i.e. in a tweet) which exist in NRC .
- **NRC-negative-feature**: contains the frequency of negatively scored words (i.e. in a tweet) which exist in NRC .

4.3.1.1.4 Slang Sentiment Lexicons

OSNs users tend to use their daily informal language when interacting online [25]. Furthermore, they use many abbreviations for faster communication and due to limited space provided for writing. Hence, the daily informal language used provides a more convenient tool for communication than a formal language does. As a result, we propose to use a sentiment lexicon designed for online slang language to collect useful information related to expressing opinions or emotions which might be missed in the formal-word lexicons. In this thesis, we use SlangSDlexicon [230]. SlangSD contains 96462 slang words labelled as positive, neutral, or negative. The three features created, based on the presence of words from our tweets in the SlangSD lexicons, are as follows:

- **SlangSD-positive-feature**: contains the frequency of positively scored words (i.e. in a tweet) which exist in SlangSD.
- **SlangSD-neutral-feature**: contains the frequency of neutral scored words (i.e. in a tweet) which exist in SlangSD .
- **SlangSD-negative-feature**: contains the frequency of negatively scored words (i.e. in a tweet) which exist in SlangSD .

4.3.1.1.5 OSNs Linguistic Hints

Beside using slang language in OSNs, the use of intensifiers like all-caps and character repetition would indicate a strong sentiment. Studies have shown that intensifiers are widely used in online conversations [124]. Therefore, we believe that they will be useful to build our classifiers. We extract four types of linguistic-hint features:

- **All-caps presence feature**: contains the presence of all-caped words, like "HORRIBLE", in a tweet.

- **All-caps frequency feature**: contains the occurrences frequency of all-caped words in a tweet.
- **Letter-repetition presence feature**: contains the presence of words with consecutive repetitive letters, like "beeest", in a tweet.
- **Letter-repetition frequency feature**: contains the occurrences frequency of f words with consecutive repetitive letters in a tweet.
- **Exclamation-mark presence feature**: contains the presence or absence of exclamation mark in a tweet. Literature states that the use of exclamation mark could indicate a strong feeling [34]. In addition, sentences ending with exclamation mark would convey emotion rather that stating a fact.
- **Exclamation-mark frequency feature**: contains the frequency of exclamation mark in a tweet. Previous works claim that consecutive use of exclamation mark would increase the attention to the feel of the opinion expressed [34].
- **Question-mark frequency feature**: contains the frequency of question mark in a tweet. Consecutive question marks are a sign of opinion intensification [143].

4.3.1.1.6 Iconic Emotion

Punctuation-based emoticons and emojis have become a ubiquitous part of OSNs culture. Users intensively use them when communicating online. Sometimes, users tend to emphasize them as they express their feelings more than words do, as well as to save space for more information to share. The latter case is especially for Twitter since it limits posts to 140-280 words. Emoticons and emojis have proven to have an important communicative role in areas like opinion expression and conversation ambiguity clarification [96]. In this thesis, we follow two approaches to extract the features related to the use of emoticons and emojis: (1) their presence/absence, (2) the sentiment they provide. For the sentiment approach, we utilized AFFIN-emoticons lexicons [159] and emoji sentiment lexicon [163].

For emoticons, the features extracted are:

- **Emoticon presence feature**: contains the presence or absence of emoticons in a tweet.
- **Emoticon-positive frequency feature**: contains the frequency of positively scored emoticons (i.e. in a tweet) which exist in AFFIN-emoticon lexicon.
- **Emoticon-negative frequency feature**: contains the frequency of negatively scored emoticons (i.e. in a tweet) which exist in AFFIN-emoticon lexicon.

For emojis, the features extracted are:

- **Emoji presence feature**: contains the presence or absence of emojis in a tweet.

- **Emoji-positive frequency feature**: contains the frequency of positively scored Emojis (i.e. in a tweet) which exist in emoji sentiment lexicon.
- **Emoji-negative frequency feature**: contains the frequency of negatively scored Emojis (i.e. in a tweet) which exist in emoji sentiment lexicon.

4.3.1.1.7 Hashtag

The presence of a hashtag in online posts gives a weight to the aspect it represents whether it is a topic, event, or emotion. It is specific to Twitter and its popularity has expanded to cover all social media arenas such as Facebook, Instagram, and Flicker. A hashtag summarizes the overall opinion of texts. Its short length nature makes the choice of its word(s) reflect stronger feeling or opinion. In the example "players didn't show their best today. # shame", the hashtagged word "shame" emphasizes a negative sentiment more than the message of the tweet itself. In this thesis, we explore hashtag as an individual type of feature. We extract two features from this section:

- **Hashtag presence feature**: contains the presence or absence of hashtags in a tweet.
- **Hashtag frequency feature**: contains the frequency of hashtags in a tweet

4.3.1.2 DFMSD Dataset Classes Features Relation

The first part of Table 4.2. shows the frequency of the top 10 words from our dataset in positive, neutral, and negative classes. We have observed that most of the top 10 words are stop words. We have also observed that the frequency of their use appear to be nearly equal among the positive, negative, and neutral classes. For example, the word "my" is used $\approx$ 17% and 16%, and 11% times in positive, negative, neutral classes respectively. Note that the frequency of the word "my" for the neutral class appear to be far less than the positive and negative classes. This is due to the class imbalance. Besides, the neutral class has the minority number of instances in comparison to the positive and negative classes. However, the percentage of its term frequency is close to the other classes. This observation follows Zipf's law that states that words with low usage frequency rank are used more often while words with high usage frequency rank are used rarely [170]. From this observation, we claim that the use of stop words to find a relationship between the features of our classes, will not be of help. So, we decided to remove the stop words. In this analysis, we have used 21641 terms, after removing the stop words. The second part of Table 4.2 shows the top 10 words among the three classes after removing the stop words. Now we can see that some of the words started to give useful information about positive class like "happy" and "love". They have much high frequency in positive class than in other classes. There are still some high frequent words that provide neutral sentiment (e.g. "day" and "just"); however, those words will not impose an importance in learning the positive class characteristics.

Table 4.2: **Top 10 term frequencies from our dataset, with and without stop words for positive, negative, and neutral classes.**

Top 10 term frequencies with stop words for DFSMD analysis			
Word	**Positive Frequency**	**Negative Frequency**	**Neutral Frequency**
the	2453	2891	1114
to	2170	1948	931
and	1286	1203	478
you	1403	902	508
in	1085	1056	473
of	1041	1045	520
is	971	1062	445
for	1167	809	478
my	1143	767	338
it	796	896	317
Top 10 Term frequencies without stop words for DFSMD analysis			
Word	**Positive Frequency**	**Negative Frequency**	**Neutral Frequency**
exam	442	341	259
just	301	423	171
day	470	266	77
like	270	344	179
love	564	51	38
today	338	181	76
time	208	219	92
tomorrow	107	357	37
new	254	133	101
happy	442	25	9

In order to find a relationship between the features in different classes, we need to decide on a metric that can capture the characteristics of words belonging to each class. By using the frequency metric only, as seen in Figure 4.2. for the positive and negative classes, we are not able to infer any meaningful relationships between the features of the classes. We have observed that most of the words fall below 600 usage frequency which makes it difficult to infer a meaningful correlation. On the other hand, very few words have high frequency from which we can infer an inverse relation between the words in two different classes. For example, high frequent words in positive classes have low frequency usage in negative class. It is important to mention that stop words were removed for the purpose of analysing our dataset (DFSMD) and were kept for the learning process.

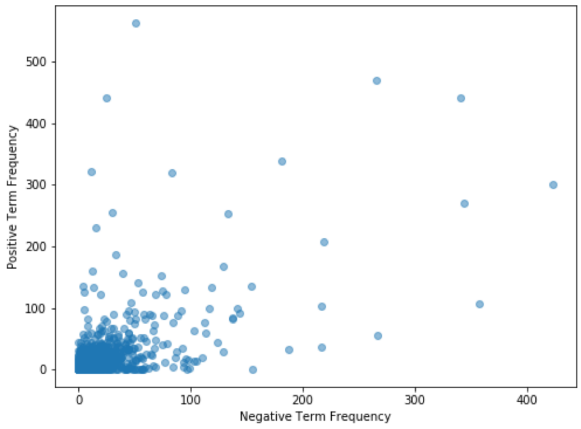


Figure 4.2: **Relationship between positive and negative classes in term of term frequency metric.**

So, the assumption here is that high frequent words that appear in a class more than in another will be useful features to learn that class. Accordingly, we have used Eq.4.2 to compute the ratio of a word belonging to a class with respect to the total frequency of the same word in all the classes. Also, we have used Eq.4.3 to calculate the ratio of a word belonging to a class with respect to the overall frequency of the same class.

$$f_{k_i}(w) = \frac{frequency_{k_i}(w)}{\sum_{c=1}^{C} frequency_{k_c}(w)} \tag{4.2}$$

$$g_{k_i}(w) = \frac{frequency_{k_i}(w)}{\sum frequency_{k_i}} \tag{4.3}$$

Eq.4.2 yields good results in cases where the frequency of words belonging to a class is very high compared to the other classes. For example, the word "fabulous" has 7 occurrences in positive class where it appears 0 times in negative and neutral classes. However, the frequency of occurrences of these words is too low to consider as features to learn the class. Therefore, it is not possible to generalize a relationship from this equation. From Eq.4.3 we could not capture useful characteristics of words to be used as a useful measure to learn distinct classes since the ratio reflects the same information as the word frequencies. Besides, we already have seen the limitation of using only frequency to find

a relationship. To overcome these limitations, we use Commulative Distributed Function (CDF) as a metric to reflect the meaning of both equations and, hence, to recognize the characteristics of important words for individual classes. CDF at value w is defined as follows:

$$CDF(w) = P(X \leq w) \tag{4.4}$$

Where X is a real random variable and P is the probability that X takes a value $\leq w$.

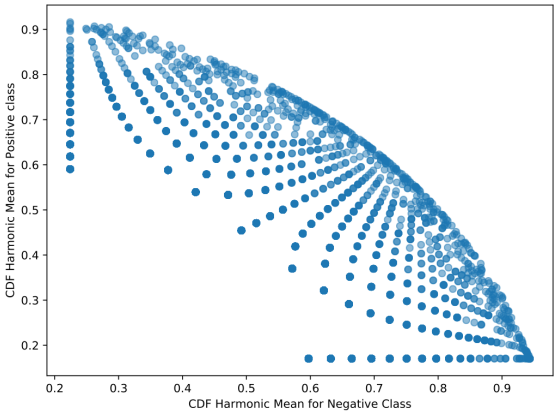
We compute CDF for both $f_k(w_i)$ and $g_k(w_i)$ in order to reflect their meanings over the words distributed among individual classes in term of accumulative manner.

Finally, we combine the CDF results for both $f_k(w_i)$ and $g_k(w_i)$ in a hope to provide a better capture of the characteristics of important words for each class. We use Harmonic Mean (Eq.4.5) due to its nature of equalizing weights given to all data points to avoid any bias towards high data points.

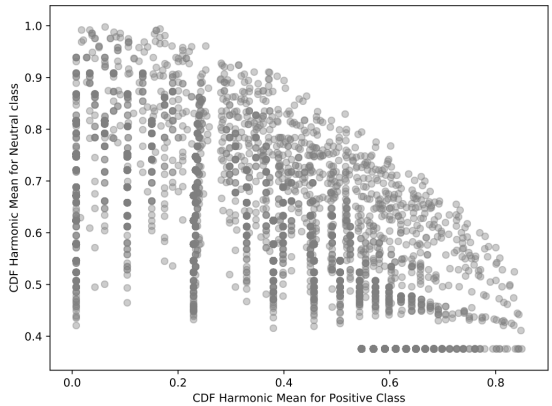
$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{4.5}$$

Where n is the size of data points and x_i is a data point.

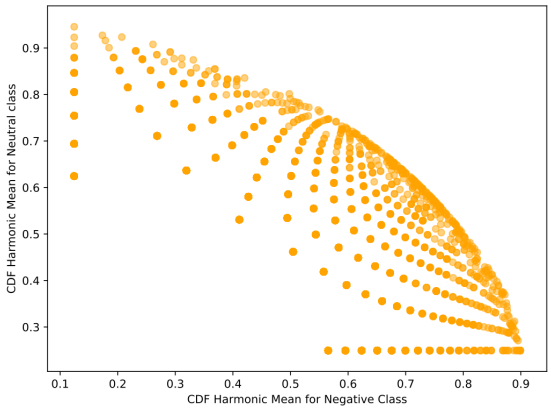
From Figure 4.3(a). we can infer that there is a relationship between words in positive and negative classes. Points with positive/negative high frequency have low negative/positive frequency. This can be seen in the data points close to the upper left corner and points close to the bottom right corner, respectively. Data points with positive CDF harmonic mean greater than 0.5 and less than 0.5 for negative CDF harmonic mean represent the important words for the positive class. The same applies to negative and neutral class as illustrated in Figure 4.3(c). From the same figure, we can see that the number of important words for the neutral class is less than that of the negative class. This is due to the class imbalance for the neutral class. On the other hand, we are not able to draw a clear relationship in the case of positive and neutral classes as seen in Figure 4.3(b). They seem to share many words with close usage frequencies. This is not surprising since neutral words tend to be closer to the positive side than to the negative one. In other words, the negative expressions tend to be strongly subjective. If we have three centroids, one for each class, the centroid of neutral class will be closer to positive centroid than to the negative one.



(a) **CDF Harmonic Mean for negative against positive.**



(b) **CDF Harmonic Mean for positive against neutral.**



(c) **CDF Harmonic Mean for negative against neutral.**

Figure 4.3: **Commutative Distribution Function (CDF) Harmonic Mean for class rate and class frequency for every pair of classes: (positive, negative), (positive, neutral), and (negative, neutral).**

The findings from this section show the quality of the dataset we proposed to use since we were able to determine the important words for the classes. Therefore, we claim the effectiveness of our word features, as they reflect an obvious association to the sentiment classes, in contributing to sentiment learning process.

4.3.2 BERT-based Deep Learning Approach

The advancement of deep neural networks has led to a decent improvement in several Natural Language Processing (NLP) tasks including sequence classifications [87]. Neural networks come with the capacity of mitigating the complexity of feature engineering and provide self-learning of data or feature representations. While memory neural networks with attention mechanism have been widely used to capture the sequential information of texts [49], CNNs [24,100] have been less popular. Comparing transfer learning in computer vision to that in NLP few years ago, transfer learning in computer vision was by far more successful in performing computer vision tasks. Earlier NLP efforts had been put to exploit previous knowledge by using textual embeddings [38,174] to avoid restarting training from scratch . Although these embeddings were trained on huge volumes of data, they still suffer from context-independence problem which means that word representations are the same regardless of their surrounding context. More recently, transformer-based language models such as BERT [112] and GBT-2 [176] have made a groundbreaking milestone in transfer learning in NLP. These models are capable of alleviating the complexity of feature engineering and overcoming the limitation of sequence-to-sequence issues (e.g., context-independence and negation issues) that are especially found when using machine learning approach. BERT has achieved state-of-the-art results in learning semantics of textual expressions for various problems including sentiment [63,204,222] and hate speech analyses [171,225,228].

BERT [112] is a multi-layer bidirectional Transformer encoder model. It uses Transformers' attention mechanism [216] to understands the inter-relationship among all words in a sentence. Transformers use an encoder to read input texts and a decoder to produce predictions of given tasks. For BERT, only the encoder is considered since its objective is to build language models. BERT model was built based on three concepts: (1) contextualized word representations, (2) transformers architecture, (3) pre-training language models on large corpus to be used for NLP task-specific fine tuning. Due to its deeply bidirectional contextualization, BERT provides a deeper sense of language contexts than single-direction models do. The contextual representation of a word considers both left and right contexts unlike single direction models that consider only the context of single direction. Two strategies are applied to learn the contextual representations: Mask Language Model (MLM) and Next Sentence Prediction (NSP). Before feeding input tokens into BERT, some percentage of the tokens are randomly masked by MLM model which then predicts the original value of masked tokens only based on the context of unmasked tokens. This way, the information about the predicted tokens is ensured not to leak to next layers. Next comes the role of NSP model where pairs of sentences (A, B) are selected from the data corpus. NSP model trains a binary classifier to predict whether the following sentence B is the actual next sentence of A. This is important to understand the

relationship between sentences and to obtain language models that have a deeper sense of a language flow and context. The resulted high-level contextualized word representations are transferable to a downstream of NLP tasks (e.g. task-specific fine turning).

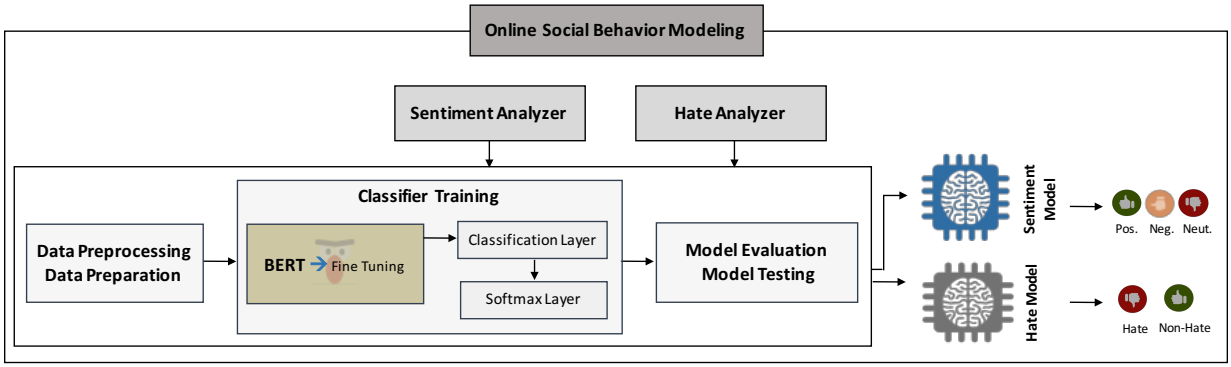


Figure 4.4: **Framework for DL-based online social behavior modeling.**

Our study focuses on two types of online social behaviors: sentiment and hate behaviors. Figure 4.4 illustrates the proposed methodology followed in order to build sentiment and hate analyzers. Supervised classification approach is adopted in the modeling of OSB. The data preprocessing is performed independently of the DEI modeling. The processed data is fed into the training component. The Classifier Training component demonstrates the proposed neural network architecture. The BERT layer consists of BERT pre-trained embeddings which are representations of words and their relation to each other in n-dimension. BERT pre-trained model is fine-tuned by training the entire BERT architecture on our datasets in order to alleviate possible biases resulted from pre-training on Wikipedia corpus [178]. BERT-base-uncased model is used in this work. It consists of twelve layers and uses 110M parameters. A feed-forward neural network layer used as a classification layer is appended to the BERT layer. The classification layer produces logits that indicate the likelihood of a tweet belonging to a class. Soft max layer is used to normalize the output logits and calculate the probability of classes. The training is conducted by back propagating the errors throughout our architecture and updating the weights of the pre-trained weights and the weights of the appended layer based on our datasets. The models are optimized using Adams optimizer. Early Stopping Approach is used to avoid over-fitting the neural network on the training data and improve the generalization of the models. Finally, we evaluate and test our models on the validation and test sets before generating the final prediction results. The prediction of sentiment analyzer is one of three classes: positive, negative or neutral sentiments, whereas the hate analyzer predicts one of two classes: hate speech or non-hate speech.

4.4 Data Preprocessing

Preprocessing data is a very essential step in machine learning in general. It prepares the resource knowledge for the machine models to learn from. High quality preprocessing ensures the quality of the learning process. The objective of preprocessing data is to remove

excess noise, which could affect the learning performance, and retain useful information. Following are the preprocessing steps we propose to use:

- **Removing Retweets**
- **Removing extra whitespaces**
- **Removing encoding symbols**: Since we use Twitter data, we might encounter cases where HTML encodings have not been converted into text. Hence, there was a need to clean this noise.
- **Removing user mention**: User mentions do not provide any opinion hints; therefore, we decided to remove them.
- **Removing URLs**: Even though URLs provide information, they might contain long texts, images, or videos. This requires different preprocessing steps. Therefore, we decided to remove them.
- **Converting text to lower case**: Treating a word that appears capitalized in the beginning of a sentence and the same word appearing lowered in the middle of a sentence would result in redundancy, hence declining performance accuracy. In addition, keeping words with upper case initials is not useful in building our sentiment model since we do not do Entity Recognition or any related tasks. We generally look for words that capture opinion, sentiment, or emotion. However, we exploit the sentiment clues that might exist in all-caps [124] words before we convert texts into lower case.
- **Expanding abbreviations**: to replace abbreviation words with sequence-of-words format. This step processes contraction words "e.g. we're", negation words (e.g. "don't"), and slang words "e.g. ppl, bro".
- **Fixing slang, negation, and repetition**: In OSNs, users tend to use abbreviations to express opinions due to the limited space. Also, they tend to use their daily language (i.e. informal or slang) and probably with the same tone they speak in real life. For example, when a strong opinion occurs, users tend to intensify words that express what they feel. "I LOOOOVE this pic" and "It isn't easy to 4gv" are example of how OSNs people tend to use short form in posts. Human eyes understand the abbreviations when encountering them because they know the origins of these abbreviations. The same analogy applies to learn our models. In our work, we apply three types of cleaners to put the words in our corpus to their original forms: (1) fixing slang: to replace slang words used in OSNs to its original terms like "bff " will be replaced by "best friend forever". (2) fixing negation: to replace abbreviated negation words to its original components. In "she isn't worried" example, we see a problem of the negation word "isn't" after tokenization; the word "t" (i.e. "not") will be meaningless. In "she isnt worried" example, the model will treat "isnt" differently than "isn't" and eventually they will be learnt as different tokens even though they are the same. Hence, this would harm our model learning. As a result, we replace

"isn't" and "isnt" by "is not". (3) fixing repetition: to remove character repetition and replace it with a single character. For example, word "niiice" will be replaced by "nice". Note that the repetition is fixed only if the repetition of consecutive characters is greater than 2. This is to ensure the originality of words that inherently consist of two consecutive characters. We believe in the importance of this step as it ensures the generality of learning sentiments of words (i.e. especially when using sentiment lexicons) regardless of their positions in sentences. It is worth to mention that we take a note of all-cap words and character repetition before the cleaning process in order to use them in feature engineering later on.

- **Removing special characters and numbers**: Numbers and most special characters do not contain sentiment insights; as a result, we decided to ignore them in this thesis. However, punctuation-based emoticons and emojis will be extracted and the presence of Twitter special symbols like hashtag(s) will be noted before removing them to be used later in the feature engineering. We believe that the hashtagged text, emoticons, and emojis provide useful information related to opinions or sentiments [1, 142]. Hence, they would improve the performance of the learning.

- **Converting iconic emotion into textual format**: to convert emojis and emoticons into textual representations [98].

- **Tokenizing text**: Text tokenization is the process of segmenting the text into meaningful words called tokens. The meaning of the whole text depends on these words. Therefore, it is an essential step in text classification as it helps capture the relations of words in text.

- **Removing stop words**: Many works in the literature proposed to drop stop words when training a textual classifier. This is because some stop words such as "is" and "be" will not drive the sentiment learning [26, 212]. However, removing stop words in the context of sentiment analysis could be problematic especially if the context is affected. For example, if "I, she, is, not" are stop words, then the sentence "I thought she is not happy" would be learnt as a positive sentiment which is not true at all. This step is applied to the traditional learning approach only; stop words are kept in when training models using deep learning method.

  Therefore, we decided to include this cleaning step in this work to investigate whether stop words removal will cause sensitivity to the sentiment performance or not.

Note that all the preprocessing steps were implemented using regular expressions NLTK, spaCy toolkit. Bert tokenizer was used for the tokenization of BERT-based model.

4.5 Experiment Design & Evaluation Protocol

4.5.1 Traditional Machine Learning Approach

We will investigate the capabilities of the state-of the art LGBM algorithm in learning a three-class sentiment using five types of features on the DFSMD dataset. We will then conduct evaluation comparisons of LGBM with SVM, Logistic Regression, Multinomial Naïve Bays, Random Forest, and Extreme Gradient Boost (XGB) using the same dataset. The objective of our experiments is to compare different algorithms in order to find the best player with the best settings of features. To achieve this, we propose to conduct four types of experiments: (1) to explore the optimal size of word features, (2) to study the sensitivity of keeping/removing stop words on the sentiment learning, (3) to study the effect of different subsets of features, (4) to investigate the role of hashtagged words and slang words in the sentiment learning from OSNs texts. The DFSMD dataset is used for training and testing and it is randomly split into 60% for training and 40% for testing.

During initial experiments, we observed that our classifiers had produced some wrong predictions on negative and neutral data samples more than on the positive samples. An interpretation of this behavior is related to the fact that the models encountered class imbalance. Previous studies [226] state that relatively balanced class distribution yields better results. By following the same approach of collecting the dataset, we did data augmentation to partially balance the negative class since it has the lowest ratio (19%). The class distributions have become 46%, 33%,21% for positive, negative, neutral classes respectively. It is valid to do data augmentations to fix the imbalance problem [110]. We did not use the sampling technique since we wanted to keep the data natural as much as possible, and to avoid creating biases in our data, as well.

For the evaluation metrics, we use accuracy, precision, recall and F-Score as they are commonly used in classification evaluation. Since our problem is a multi-class classification and our dataset is imbalanced, we attempt to use micro average F-Score for the evaluation. Micro average F-Score is computed by aggregating the contributions from all the classes instead of averaging individual contribution for each class like in macro-average F-Score. The accuracy is defined as the ratio of the correct predicted samplers to the total number of samples in a test set. Precision, recall and F-Score give a better view of model performance than accuracy alone does. They are calculated as illustrated in the following equations:

$$\begin{aligned} Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \\ F - Score &= \frac{2 \times P \times R}{P + R} \end{aligned} \tag{4.6}$$

4.5.2 Deep Learning Approach

Sentiment and hate are the two social behaviors studied in this work. We evaluated the performance of BERT for sequence classification algorithm using DFSMD dataset for sentiment learning and Hate dataset for hatespeech learning. We then conducted evaluation comparisons of the BERT-based models with other algorithms for sequence classification including LSTM, biLSTM, CNN-LSTM, and CNN-biLSTM, using the same datasets. Figure 4.5 illustrates the architectures that we have used in our experiments.

For BERT pre-trained model, BERT-base-uncased was used. It consists of 12 blocks of transformers, 768 hidden layers, 12 attention heads, 110M parameters. During the fintuning process, the learning rate was set to 2e-5, epochs were set to 6, and the batch size was 16.

GloVe [174], a pre-trained word-embedding, was used to train LSTM, biLSTM, CNN-LSTM, and CNN-biLSTM models. The data preprocessing steps and experiment setups were the same as those of the BERT-based models.

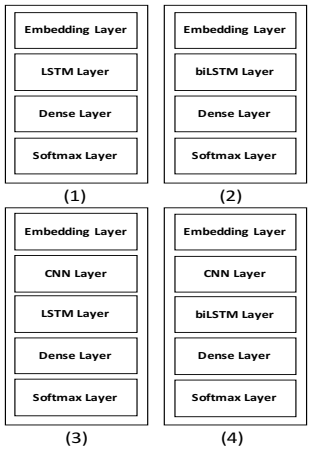


Figure 4.5: **The neural network architectures used for comparison with our BERT-based models. (1) for LSTM, (2) for biLSTM, (3) for CNN-LSTM, and (4) for CNN-biLSTM**

The DFSMD and Hate datasets were used for training, validation, and testing sentiment and hate models, respectively. The datasets were randomly split into 70% for training, 15% for validation and 15% for testing. We used accuracy, precision, recall, and F-Score, commonly used for classification evaluation, as evaluation metrics. Precision, recall and F-Score give a better view of model performance than accuracy alone does.

4.6 Results and Analysis

4.6.1 Traditional Machine Learning Approach

4.6.1.1 Size of Word Features

In this experiment we will investigate the optimal number of word features to be used in building our sentiment classifier. Based on the findings from Section 4.3.1.2, we claim that the words extracted from our dataset have distinguishing characteristics for classes especially for the positive and negative classes. From our dataset, we have extracted 21310 vocabulary words including stop words and 21030 words excluding stop words, which is a large number. Therefore, we need to find out a reasonable number of words to use as features to train our sentiment classifiers. We choose to examine word sizes of 1000, 2000, 3000, 4000, and 5000 on six classifiers using BOW and n-gram features. Note that the sizes represent the maximum number of words based on their term frequency TF. For n-gram models, we examine uni-gram, uni-bi gram, and uni-bi-tri gram models. While investigating the number of features, we explore the impact of removing and keeping stop words on the sentiment learning process. As a result, our experiments are conducted on six different combinations of BOW and n-gram models with and without stop words. The results are illustrated in Figure 4.6.

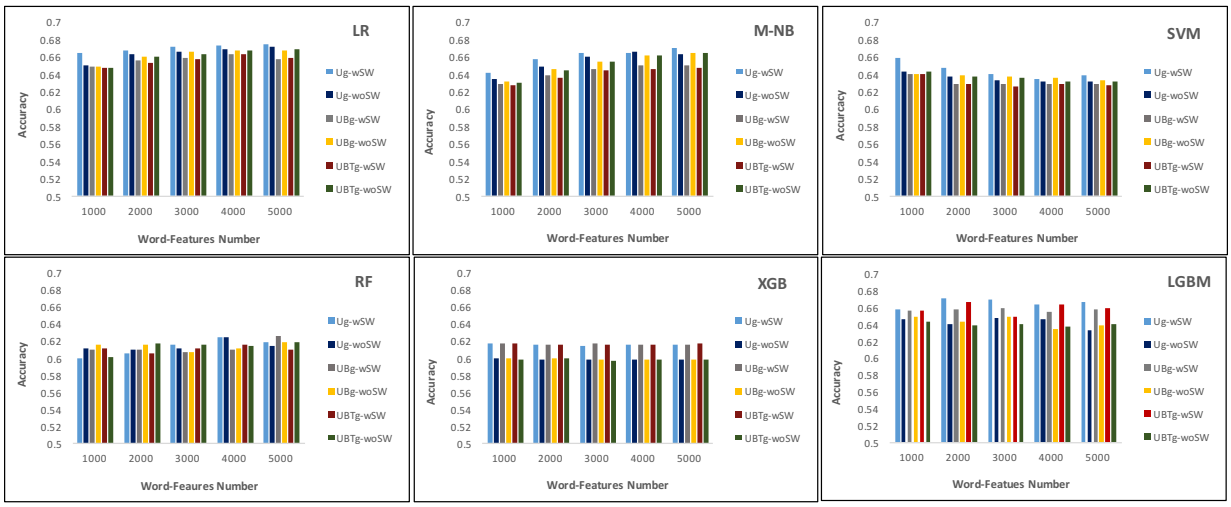


Figure 4.6: **Performance of TF-BOW and n-gram with/without stop words using different vocabulary sizes evaluated by accuracy. Ug, Bg, and Tg stand for uni-gram, bi-gram, and tri-gram, respectively. wSW, woSW stand for with stop words and without stop words, respectively.**

From the results, we see that the best performance occurs with uni-gram model when keeping stop words. This finding should not be surprising since we use short texts (i.e. tweets) of length average of $\approx$ 76 tokens. In other words, stop words seem to be of importance in learning sentiment in this work. On the other hand, uni-bi gram and uni-bi-tri gram models work better when removing stop words. However, their overall performance did not exceed uni-gram model when stop words were kept. As a result, we consider uni-gram model with keeping stop words in our sentiment classification.

Table 4.3: **F-Score, precision, and recall for positive, negative, and neutral classes when using all proposed features.**

	LR			MNB			SVM			RF			XGB			LGBM		
	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall
Positive	**0.79**	0.75	0.83	0.78	0.71	0.88	0.78	0.75	0.81	0.75	0.67	0.84	0.77	0.73	0.82	**0.79**	0.77	0.81
Negative	0.76	0.71	0.8	0.73	0.67	0.82	0.76	0.73	0.8	0.7	0.68	0.72	0.74	0.71	0.77	**0.77**	0.74	0.8
Neural	0.36	0.5	0.28	0.14	0.59	0.08	0.4	0.48	0.34	0.2	0.39	0.14	0.33	0.45	0.26	**0.41**	0.48	0.35

According to the results depicted in Figure 4.6, the optimal size window seems to differ between *uni*, *bi*, *tri* grams when keeping or removing stop words, among the six classifiers.

For example, the logistic regression (LR) and Multinomial Naïve Bays (MNB) models have the best performance at words size of 5000 when using *uni*-gram with stop words. Words size of 1000 shows to be the best with Support Vector Machine (SVM) model when using *uni*-gram and keeping stop words. Random forest (RF) classifier shows best similar results when using 4000 and 5000 words with and without stop words on *uni*-gram model and with stop words on *uni*-bi gram, respectively. For Gradient Boost (XGB) model, the best performance appears to be similar among all the size windows on uni-gram including and excluding stop words and $(uni - bi - tri)$-gram including stop words. Even though $(uni - bi - tri)$-gram model with stop words seems to have the best performance, the difference in performance accuracies is almost negligible. Finally, *uni*-gram with stop words wins at 2000 size window when training Light Gradient Boosting Machine (LGBM) classifier.

Since we consider uni-gram with stop words model as it yields the best performance for all the classifiers, we consider the size windows where each classifier works the best. Then, we average all the sizes and use the average as the maximum number of word features to train our classifiers for the rest of the experiments. The average of maximum number of features used in this work is 3000 words.

4.6.1.2 Cross Features Subsets

Table 4.4. shows the results of six sentiment classifiers of fifteen cross combinations of our proposed feature types explained in Section.4.3.1.1. We split the experiments into four stages: (1) examining individual feature types, (2) examining formal lexicon (FL) feature with different combinations of the rest of the feature types excluding bag-of-words (BOW) feature, (3) examining BOW feature with different combinations of the rest of the feature types. We decided to examine FL and BOW as main features in stage 2 and 3 due to their highest contributions in sentiment learning among the other feature

types. From the same table, we can see that using BOW features only yielded the highest performance accuracy and F-Score among all the classifiers, followed by FL feature. While their accuracy scored above 60% and average F-Score (i.e. micro) is above 0.60, the rest of feature types scored less than 50% accuracy and 0.50 micro average F-Score when used individually. Linguistic hint features (Lngs) seem to perform better than iconic emotion (Emt-Emj), slang lexicon (SL), and hashtag that comes at the lowest performance rank. This finding shows an evidence of the association between our word features and sentiment classes. Hence, our classifiers are able to properly learn the distinguishing characteristics for each class. In addition, the result of using FL feature shows that a high percentage of our dataset vocabulary is contained in the lexicons, which also shows the distinguishing characteristics of our word features. Moreover, we have observed that the more formal lexicons we add, the better vocabulary coverage we obtain, which in turn improves the sentiment learning. In this work, we have decided to use two formal lexicons as discussed in Section 4.3.1.1.

Table 4.4: **Performance of fifteen cross features sets on six classifiers evaluated by accuracy and F-Score.**

	LR			M-NB			SVM			RF			XGB			LGBM		
Features	**Accuracy**	**F-MacAve**	**F-MicAve**	**Accuracy**	**F-MacAve**	**F-MicAve**	**Accuracy**	**F-MacAve**	**F-MicAve**	**Accuracy**	**F-MacAve**	**F-MicAve**	**Accuracy**	**F-MacAve**	**F-MicAve**	**Accuracy**	**F-MacAve**	**F-MicAve**
BOW	67.65	0.611	0.68	66.18	0.611	0.672	64.68	0.6	0.654	61.68	0.531	0.604	60.51	0.465	0.605	65.86	0.573	0.656
Formal Lexicon (FL)	62.32	0.455	0.62	61.7	0.448	0.614	62.3	0.454	0.62	61.89	0.48	0.625	62.46	0.48	0.624	61.96	0.479	0.624
Slang Lexicon (SL)	46.26	0.21	0.46	46.26	0.21	0.46	46.26	0.21	0.46	48.53	0.342	0.485	50.07	0.35	0.492	48.41	0.348	0.492
Linguistic (Lngs)	48.84	0.355	0.499	47.84	0.274	0.483	48.84	0.355	0.499	48.78	0.355	0.5	50.28	0.355	0.498	48.83	0.355	0.499
Emoticon-Emoji (Emt-Emj)	47.2	0.246	0.475	46.62	0.25	0.47	47.22	0.246	0.475	47.2	0.246	0.475	47.58	0.247	0.475	47.2	0.247	0.475
Hashtag	45.88	0.246	0.475	45.88	0.25	0.47	45.88	0.246	0.475	45.88	0.246	0.475	46.3	0.247	0.475	45.88	0.247	0.46
FL+Lngs	62.27	0.478	0.622	61.16	0.453	0.612	62.35	0.477	0.622	60.96	0.486	0.605	64.55	0.502	0.622	62.49	0.505	0.619
FL+Lngs+SL	63.23	0.497	0.634	61.87	0.468	0.625	63.11	0.491	0.632	58.76	0.519	0.594	65.09	0.512	0.636	63.04	0.527	0.634
FL+Lngs+SL+Emt-Emj	64.8	0.541	0.648	63.85	0.472	0.637	64.94	0.512	0.647	60.47	0.546	0.604	66.19	0.561	0.651	64.79	0.57	0.648
FL+Lngs+SL+Emt-Emj+Hashtag	65.79	0.57	0.653	64.68	0.495	0.649	65.29	0.54	0.655	61.78	0.543	0.605	66.58	0.584	0.655	65.68	0.584	0.648
BOW+FL	69.01	0.582	0.682	68.5	0.636	0.666	65.03	0.63	0.685	65.22	0.571	0.646	68.53	0.591	0.66	69.1	0.63	0.688
BOW+FL+SL	68.51	0.59	0.687	68.5	0.517	0.669	65.49	0.617	0.686	64.96	0.54	0.643	68.5	0.542	0.661	68.65	0.603	0.68
BOW+FL+SL+Lngs	69.03	0.626	0.69	68.75	0.627	0.688	66.1	0.611	0.661	65.23	0.559	0.652	69.81	0.566	0.673	69.15	0.616	0.692
BOW+FL+SL+Lngs+Emt-Emj	70.13	**0.647**	0.704	69.19	**0.637**	**0.697**	66.25	0.629	0.674	66.2	**0.573**	**0.665**	70.46	0.591	0.684	70.96	0.647	0.707
All	**71.1**	0.635	**0.707**	**69.74**	0.552	0.69	**67.86**	**0.648**	**0.707**	**66.25**	0.551	0.656	**71.01**	**0.612**	**0.687**	**71.79**	**0.654**	**0.712**

Despite the low performance that slang, linguistic, iconic emotion, and hashtag features show when used individually, they still provide some sentiment signs that would enrich the sentiment learning. This can be seen from the results of stages 2 and 3 in two observations. First, their performance accuracy is approximately close to 50% and micro average F-Score approximately close to 0.50. Even though this percentage is not high enough, it still shows that these features might carry some sentiment information that could be used as supplementary features along with BOW and formal lexicon features. Second, the more features we combine the better learning performance we obtain. In stage 2, LGBM model trained solely on formal lexicon features yields performance accuracy of $\approx$ 62% (micro average F-Score of 0.62). When combining formal lexicon with linguistic feature the performance stays the same; however, when adding the slang feature, the accuracy and F-Score slightly improved to 63% and 0.63, respectively. The performance accuracy improves to $\approx$ 64.8% (micro average F-Scoreof $\approx$ 0.65) with the addition of the iconic emotion feature. When combining formal lexicon features with the rest of the features, the performance improved by $\approx$ 3% more than when the model was trained on formal lexicon alone. The other five classifiers show similar results.

In stage 3, we consider BOW as a base feature and combine it with different subsets of features. Combining the two most contributing features has shown to boost the performance by $\approx$ 7% more than when LGBM, LR, MNB, XGB models were trained on only FL feature. The same classifiers trained on BOW only improved by $\approx$ 2-3% when combining

Table 4.5: **F-Score, precision, and recall for positive, negative, and neutral classes when using slang lexicon feature only.**

	LR			M-NB			SVM			RF			XGB			LGBM		
	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall
Positive	**0.63**	0.46	**1**	**0.63**	0.46	**1**	**0.63**	0.46	**1**	0.61	0.497	0.789	0.608	**0.502**	0.77	0.609	0.501	0.776
Negative	0	0	0	0	0	0	0	0	0	0.407	0.463	0.364	**0.44**	0.47	**0.413**	0.435	**0.472**	0.404
Neutral	0	0	0	0	0	0	0	0	0	**0.008**	**0.106**	**0.004**	0.002	0.091	0.001	0	0	0

BOW and FL features. As illustrated in Table 4.4., combining BOW with all the feature types yields the best performance among all the classifiers and all different feature subsets. Compared to the performance of models when trained only using BOW, accuracy of LGBM model has improved from 65.86% to 71.79% when combining all the features together. The same results apply to LR, MNB, SVM, RF, and XGB when all the features are combined. Even though slang lexicon feature and linguistic feature show to have sentiment signals, they do not seem to add value when they are not combined together in a feature subset. This can be seen in two cases: (1) when linguistic feature is combined with formal lexicon feature, (2) when slang lexicon feature is combined with BOW and formal lexicon features. However, when slang lexicon and linguistic features are combined together, they seem to add a little value to the learning process. This shows that the tweets might contain slang words written with intensifiers such as "WTH" and "looool". Moreover, precision and recall of the linguistic feature for neutral class is zero among all the classifiers. This is not surprising because intensifiers like all-caps and repetitive-characters are usually used when there is an opinion that needs to be stressed on. On the other hand, the precision and recall for positive and negative class are quite reasonable and reflect the performance of the classifiers and the distribution of the positive and negative classes. In addition, internet language of emoticons, emojis, and hashtags have shown to have an impact on the sentiment learning among all the classifiers. Emoticons and emojis have proven to enhance the classification accuracy from 63% to $\approx$ 64.8% in LGBM model when combined with formal lexicon, linguistic, and slang lexicon features. Furthermore, adding the hashtag feature to the previous combination has shown to increase the LGBM model accuracy to $\approx$ 65.7%. This result is consistent when combining emoticons and emojis with BOW, formal lexicon, slang, and linguistic feature while training LGBM model. The performance accuracy has improved from 69% (i.e. using BOW and FL) to $\approx$ 71%. It further improved after adding the hashtag feature to reach $\approx$ 71.8%. In terms of precision and recall for neutral class when training using only emoticons and emojis features, they are shown to have zero values. Again this is not surprising; emoticons and emojis are usually used for emotional expression such as "smiley face" and "angry face". Based on this finding, we have observed that people use emoticons and emojis to express subjective opinion rather than objective ones.

LGBM model shows the best learning performances in term of accuracy and F-Score, followed by LR. Further, MNB and XGB models are shown to be strong competitors followed by SVM model. However, the only problem with XGB is that it is too slow to converge in comparison with the other five algorithms. This result is consistent with the findings of LGBM's authors [111] where XGB has shown a noticeably slower convergence rate and less learning performance than LGBM. RF classifier is shown to have the weakest

performance among the rest. This shows the efficiency of gradient boosting algorithms compared to the bagging technique adopted by RF. Gradient boosting techniques used in LGBM and XGB make use of trees with fewer yet better quality splits instead of growing trees to their maximum extent. Also, the techniques used in LGBM shows their strength in dealing with difficult cases (i.e. neutral class in our case since it is the minority class). LGBM has learnt neutral class with F-Score value of $\approx$ 0.41 in comparison with 0.4, 0.36, 0.32, 0.20, 0.13, for SVM, LR, XGB, RF, MNB (shown in Table 4.3.). As a result, we attempt to use LGBM, with all the proposed features to build our general sentiment classifier using our dataset.

The main limitation of the proposed dataset is the data imbalance, especially for the neutral class. Despite this fact, our models give good performance results even on neutral class for some classifiers. Positive and negative classes scored the highest learning performance values. For some of the models, an acceptable performance was yielded for the neutral class. Table 4.3. illustrates our models performances across the three classes, positive, negative, neutral. From this result, we have observed that our dataset contains separating characteristics for the three classes in which the classifiers could successfully learn from. Further, the proposed OSN-specific features have proven their supplementary effect on the sentiment learning. Another limitation is that this study focused on predicting explicit sentiments from social media texts but was not designed to predict implicit sentiments contained in sarcastic texts. Despite this limitation, the iconic features (i.e. emojis and emoticons) could assist in recognizing sentiments in this case.

4.6.1.3 Cross-Domain Sentiments

In this section, we will present two experimental scenarios for cross-domain sentiment prediction: (1) examining our general sentiment classifier (i.e LGBM) on datasets of two domains: movie reviews and sports, (2) examining domain-specific sentiment models (i.e. movie reviews and sports) on our general sentiment dataset (DFSMD).

We trained two domain-specific LGBM models; one for IBDM movie reviews and another for sports (CL'16-'17) tweets. We used the same exact experimental settings and features (i.e. all proposed features combined) that we used to train our general LGBM model. We split the data into 60% for training and 40% for testing. The LGBM sentiment model trained on IMDB two-class dataset performs well at accuracy score of $\approx$ 86% (micro Ave. F-Score of 0.86), whereas LGBM classifier trained on three-class sports CL'16-'17 dataset yields an acceptable learning performance at $\approx$ 57% of accuracy (micro Ave. F-Score of 0.57)).

We conducted our experiments on three datasets as shown in Table 4.6. For IMDB and CL datasets, we used a subset of their instances for the purpose of our evaluation.

Table 4.7 presents the results of adapting our general domain sentiment model (LGBM)to domain-specific sentiment datasets. It can be seen that it is effective to adapt general sentiment modelling to domain-specific sentiment analysis. Our LGBM general model shows a good sentiment prediction performance on both movie reviews and sports tweets, for positive and negative classes. Our LGBM model was able to recall 51%, 67% (precision

Table 4.6: **Details of three sentiment datasets used in cross-domain experiments. DFSMD is used for general domain sentiment, IMDB is used for movie reviews domain sentiment, and CL'16-'17 is used for sport domain (Champions League) sentiment.**

	DFSMD [7]	**IMDB** [133]	**CL-'16-'17** [8]
Source	Twitter	IMDB	Twitter
Text Length Ave	81	1270	82
Instances	14,488	25,000	14,000
No. of Positive	6683	11500	5150
No. of Negative	4822	11500	4275
No. of Neutral	2983	-	5442

of 0.60, 0.71) of the positive, negative instances of IMDB dataset. We observed that the positive recall is lower than the negative one. This could be due to the fact that review texts have special dictionary and sentence patterns that do not necessarily exist in general conversations. However, our general LGBM model could correctly recognize $> 50\%$ of the positive class with 62% precision. The result is actually promising since our LGBM was trained on short texts to learn three classes, while IMDB reviews are of long texts (see Table 4.6) and only consist of two classes. This finding indicates that people generally use the common words and phrases to express opinions, regradless of the texts lenght. Again, this shows the quality of our dataset (DFSMD), in terms of data contents and annotation, for sentiment classification. In addition, our LGBM classifier performs even better on the sports CL'16-'17 dataset; it could successfully recall 71%, 64% (precision of 0.74, 0.71) of positive, negative instances. This is not surprising as the CL'16-'17 dataset consists of short text tweets. We can see that the recall of our general LGBM model is slightly lower in the sports domain than that of the movies reviews. This is due to the fact that sports domain reserves a special language where many terms and phrases indicate sentiments contrary to those of general original sentiments [7].

Table 4.7: **Sentiment performance of our general sentiment LGBM model on two domain-specific datasets: IMDB movie reviews and CL'16-'17 tweets.**

	Positive			**Negative**		
	Precision	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**
General DFSMD -> IMDB	0.62	0.51	0.56	0.6	**0.67**	0.63
General DFSMD -> CL'16-'17	**0.74**	**0.71**	**0.73**	**0.71**	0.64	**0.67**

Generalizing domain-specific sentiment modelling is shown to be less effective than adapting general sentiment modelling to domain-specific analysis. From Table 4.8, we can observe that domain-specific LGBM models could not properly learn the negative instances of the general dataset even though they yield good learning performance on the positive instances. This implies that domain-specific data introduces learning confusion to general sentiment [7] as some terms might indicate negative sentiments in a specific domain but

Table 4.8: **Sentiment performance of two domain-specific LGBM models trained individually on IMDB and CL'16-'17 datasets. The sentiment performance was evaluated on our general sentiment dataset DFSMD.**

	Positive			Negative		
	Precision	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**
IMDB -> General DFSMD	0.69	**0.94**	**0.79**	**0.82**	0.41	0.54
CL'16-'17 -> General DFSMD	**0.79**	0.58	0.67	0.81	**0.43**	**0.56**

positive sentiments in general domain.

4.6.2 Deep Learning Approach

The results of modeling two online social behaviors, sentiment and hate, are demonstrated and discussed in this section. Following, a detailed analysis of the sentiment and hate online behaviors during the pandemic, is provided for the duration between December 2019 and November 2020 for both Canada and USA. The analysis presents two views: temporal and topic-based analysis.

4.6.2.1 Hate Behavior

Table 4.9: **The performance of five deep learning algorithms for sequence classification for hate classification in terms of accuracy, precision, and recall.**

	LSTM			biLSTM			CNN-LSTM			CNN-biLSTM			BERT		
	Accuracy	85		Accuracy	85		Accuracy	85		Accuracy	85		Accuracy	86	
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
Normal	0.83	0.9	0.87	0.85	0.88	0.86	0.84	0.9	0.87	0.86	0.86	0.86	0.88	0.84	0.86
Hate	0.87	0.79	0.83	0.85	0.81	0.83	0.87	0.8	0.83	0.84	0.84	0.84	0.85	0.89	0.87

Table 4.9 summarizes the learning performance of four algorithms for sequence classification using pre-trained GLOVE embeddings as features, and compares them with our proposed BERT-based classifier using its pre-trained embedding as features. It is important to mention that the embedding features were fine-tuned during the training. In terms of accuracy, it is shown that BERT-based classifier performs the best in detecting normal and hate contents in social messages compared to the other four algorithms. The other four algorithms show equal performance in terms of accuracy. However, LSTM and biLSTM classifiers show a bias towards learning the majority class (i.e. normal class). This can be seen in the relatively high variance between the recall values of normal and hate classes with the normal class value being the higher. Also, the lower value of normal precision and hate recall explains that the LSTM and biLSTM models have over learned the majority class (i.e. normal), and hence started to introduce a degree of confusion in correctly classifying hate class (i.e. the minority). Adding a layer of CNN to LSTM and biLSTM has helped solve the issue of bias learning and improve the overall learning performance in detecting

hate speech in texts. Combining CNN as a mechanism to find important features, and bidirectional learning mechanism has shown to enhance the hate learning process, more than when combining CNN with a single-direction LSTM. The variance between the recall values of normal and hate classes has decreased while still maintaining high scores of recall and precision for both classes. In comparison with CNN-biLSTM classifiers, BERT calssifier has shown more robust capabilities in tackling the issue of bias learning towards the majority class. The attention and bidirectional mechanisms adopted by BERT algorithm have proved their effectiveness in improving the quality of learning the two classes, more than CNN and bidirectional mechanisms could achieve. From Table 4.9, we can see that the hate F-Score of BERT model has improved from 0.84 (of CNN-biLSTM) to 0.87 and precision scores for both classes have improved from 0.86 and 0.84 (of CNN-biLSTM) to 0.88 and 0.85, respectively.

From the results, we can see that the attention and bidirectional learning mechanisms adopted by BERT show more efficiency in textual sequence classification than CNN combined with bidirectional learning mechanisms do.

4.6.2.1.1 English Case Study: Hate Behavior Analysis during COVID-19 Pandemic in USA and Canada

We have utilized our BERT-based hate classifier to analyze the hate speech behavior during COVID-19 pandemic in North America. The details of COVID-19 data collection for both USA and Canada can be found in Chapter 7. Figure 4.7 shows the overall hate behavior detected in Canada and USA during three periods of the COVID-19 pandemic. We see that hate behavior in USA is slightly higher than it is in Canada.

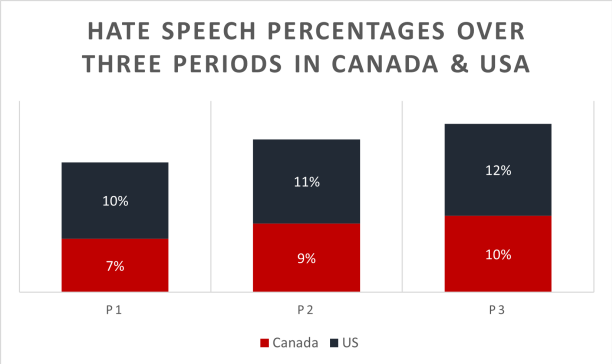


Figure 4.7: **Percentages of hate behavior in Canada and USA during periods 1 (Dec 2019 - Apr 2020), 2 (May 2020 - Aug 2020), and 3 (Sep 2020 - Nov 2002) of COVID-19 pandemic.**

In Figures 4.8 and 4.9, an exploratory analysis of hate behavior in both Canada and USA is demonstrated. We provide a detailed analysis and interpretation from two views: temporal analytic view and topic-based analytic view.

Figure 4.8 illustrates online social hate behavior at the very early signs of the virus and during the pandemic (December 2019 - November 2020). Generally speaking, online

hate behavior seems to have been lower in Canada than in USA except for the month of December 2019; the people in Canada seem to have been more upset about the virus. However, the number of COVID-related tweets during this month was very low (154 tweets) compared to the number of tweets afterwords (4M+ tweets). In USA, hate behavior marked the lowest in December 2019, and that was when the news started to talk about the novel Coronavirus just before the outbreak of the pandemic. Then, the hate behavior started to increase (in USA) till it hit its highest in February 2020 which marked the beginning of the pandemic. It also increased during February 2020 in Canada. By this time, the corona virus had been all over the news and people started to panic. Surprisingly, during the start of the quarantine and lockdown (i.e. March), the hate behavior decreased in both Canada and USA. Later from April till May 2020, it slightly increased in USA during April and then it decreased again in May. In Canada, it almost flattened out after March. From June till November, it can be seen that hate behavior was higher than it was before May especially for USA; two spikes were found in the months of July and October for both Canada and USA as seen in Figure 4.8.

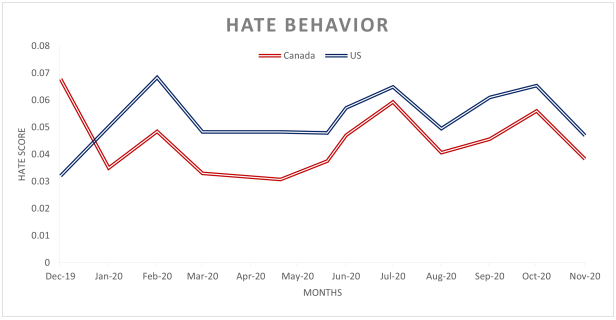


Figure 4.8: **Temporal comparisons of hate behavior over twelve months between Canada and USA before and during COVID-19 pandemic.**

Out of the twelve topics discussed during period 1 (Dec 2019 - April 2020) of the pandemic in North America (Figures 4.9(a) and 4.9(b)), we have noticed that Canada has a single hate spike (i.e. topic of "Trump&China") while USA has two hate spikes (i.e. topics of "Social Distancing" and "Trump&China"). The inferred topics for Canada and USA are listed in Table 7.7 and Table 7.8, respectively.

"Trump&China" topic in Canada and USA focus mainly on Trump's blaming China for the COVID virus outbreak. On this topic, USA shows higher hate score than Canada does. From our extracted topic phrases, we can see that among the mostly used phrases in this topic are: "lie trump", "trump blames china", "trump wartime president", "president trump shame", and "president trump beat china". The second highest spike in USA is on the topic related to social distancing and wearing mask. Phrases like "social sick", "social crazy", "fear mask", "stupid trump" provide a hint that people were not happy with social distancing and the policy of wearing masks. Again, we see hate content in topic "Face-Masks&Food-Stores" that talks about wearing masks while shopping at stores. In Canada, We observe a less hate behavior for wearing mask. Apparently, wearing masks policy has an association with hate content during the pandemic. Moreover, People in USA did not show hate speech in the topic of "Community-Health Support" While in Canada we detect

a slight increase in the hate behavior on the "community support" topic (i.e. as compared to USA). A sample of tweets related to this topic (i.e. in Canada) is given below:

- "don't forget, he has a habit of repurposing emergency funds to his wall. Watch for the noise to start about immigrants carrying the virus and showing up at the southern border".
- "Child care staff who are overworked and grossly underpaid will once again be left holding the bag. Close childcare centers, too!".
- "How stupid you guys looked when @jkenney announced medical emergency in Alberta. Please hold your horses now and see why it was important to do emergency you idiots".

Covid-19 death toll topic did not show a sign of hate speech in Canada while the hate speech on the same topic increased in USA. The tweets related to this topic were mostly news. From this, we observe that the Canadian news have less sharp reporting tone than that of the American's. Quarantine and staying home topic shows a very low hate behavior in both Canada and USA. Actually, our topic model and interpreter showed that people enjoyed being quarantined; keywords like "song", "movie", "fun", "dance", "dog", "walk", "laugh", "stayhome", "staysafe", are indicators that the quarantine could have been associated with spending good time while staying home. Cancelling sport game events, a topic talked about during the pandemic, seems to have upset people in USA more than it did with people in Canada. The effect of the pandemic on economy was also discussed in Canada and USA. People in both countries showed low hate behavior. "Pandemic-State in Quebec" topic in Canada was related to the home care in Quebec province. We observe a slight increase in the hate behavior in this topic and this might be associated with the high number of cases that hit Montreal city particularly.

It is observed that people were more adapted to the quarantine during period 2 (May 2020 - Aug 2020) more than they were during period 1 of the pandemic; topics were mostly discussing songs, movies, tv shows, birthday, hair, and food as seen in Figures 4.9(c) and 4.9(d). These topics have mostly shown low hate behavior compared to topics directly related to the pandemic such as "Trump&China" and "Sympathy Attitude" where people have expressed blames and resentments as a result to friends, family or jobs loss for example. "OMG! THATS SO, SO HORRIBLE.. I'M STUNNED. WTF KIND OF WORLD IS THIS? SO MANY CONDOLENCES TO ALL FAMILY, FRIENDS AND THE MANY ANIMALS LEROY TOUCHED SUCH A WASTEFUL, STUPID LOSS" and "I lost around 12 friends to the virus be it online friends, lose them it's painful. You feel anger, sadness; great loss, you go through the stages of grief. Trump would be acting way differently if he lost friends?" are examples of how resented people were. Similarly, "Hair" topic has shown a slight sign of hate speech as a result of barber shops being closed (e.g. "i need someone to get rid of the hair on my scalp i hate it", "I hate his hair").

The curve in Figure 4.9(e) illustrates that the hate behavior in Canada is more relaxed than it is in USA (Figure 4.9(f)). Topics inferred during period 3 (Sep 2020 - Nov 2020), from USA, reflect the political situation (i.e. elections) in the area. five out of eleven topics are related to Trump, Biden, and election with Trump related topics representing the highest hate signs. Similar behavior is found in "Trump&Biden Election" topic in

Canada as well. Interestingly, community support related topics, in Canada, have shown lower hate scores in period 2 and 3 compared to its score during period 1.

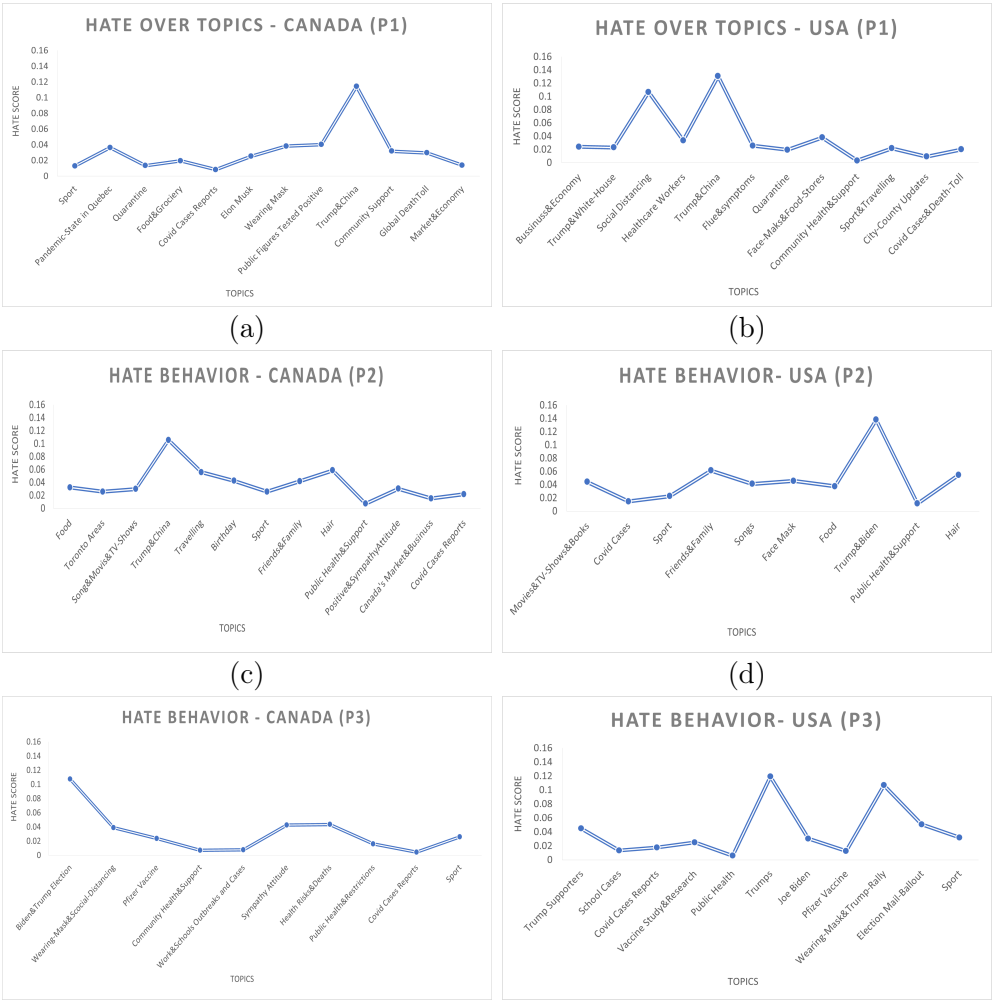


Figure 4.9: **Comparisons of Hate behavior between Canada and USA over topics inferred during periods 1 & 2 & 3 of COVID-19 pandemic. P1 (Dec 2019 - Apr 2020) is depicted in (a) and (b), P2 (May 2020 - Aug 2020) is dipected in (c) and (d), P3 (Sep 2020 - Nov 2020) is depicted in (e) and (f).**

4.6.2.2 Sentiment Behavior

Table 4.10: **The performance of five deep learning algorithms for sequence classification for sentiment classification in terms of accuracy, precision, and recall.**

	LSTM			biLSTM			CNN-LSTM			CNN-biLSTM			BERT		
	Accuracy	71		Accuracy	71		Accuracy	70		Accuracy	71		Accuracy	75	
	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score	Precision	Recall	F-Score
Positive	0.79	0.79	0.79	0.73	0.86	0.79	0.78	0.78	0.78	0.78	0.8	0.79	0.81	0.81	0.81
Negative	0.72	0.8	0.76	0.74	0.76	0.75	0.73	0.77	0.75	0.73	0.79	0.76	0.81	0.83	0.82
Neutral	0.46	0.38	0.41	0.49	0.28	0.36	0.46	0.43	0.44	0.48	0.39	0.43	0.5	0.47	0.48

In Table 4.10, we report the performances of five deep models for sentiment classification. The first four models were trained using pre-trained GLOVE embeddings as features and BERT-based model was trained using its pre-trained embedding as features. It is important to mention that the embedding features were fine-tuned during the training.

Table 4.10 shows that BERT-based classifier outperforms LSTM, biLSTM, CNN-LSTM, and CNN-biLSTM in sentiment learning. BERT model yields by far the best learning results for the three classes with F-Scores of 0.81, 0.82, 0.48 for positive, negative, and neutral, respectively. The nature of BERT algorithm, that is using attention and bidirectional learning mechanisms together, allows it to boost the learning performance across the three classes in the presence of imbalance class distribution. BERT algorithm has improved the overall learning of majority class (i.e. positive) by $\approx 2\%$ and by $\approx 8\%$, $\approx\%12$ for negative class and neutral class (i.e. minority class). The overall learning improvement is evaluated in terms of F-Score metric. It is strongly able to distinguish between classes especially for positive and negative. Both precision and recall scores for both classes are very high. Neutral class has been always challenging in sentiment classification [19]. However, BERT classifier was able to correctly recall 47% of the neutral instances at 50% of precision. The corresponding confusion matrix shows that the BERT model predicts 31% of neutral class as positive and 21% of neutral class as negative. This is not surprising since negative expressions tend to be strongly subjective. Therefore, positive expressions would be closer to neutral than negative sentiment [19].

We have observed that biLSTM introduces bias in learning the majority class. The positive recall score is higher compared to the negative and neutral recall scores, which makes the variance between them high as well. The corresponding confusion matrix shows that the model predicted 15% of the negative class as positive and 8% of the negative class as neutral. Similarly, with neutral class, 46% was predicted as positive and 25% as negative. We have observed the same behavior in biLSTM when modeling the hate behavior (Table 4.9).

Using CNN as a filtering mechanism along with biLSTM has demonstrated good performance in reducing the sensitivity to class imbalance shown by biLSTM and in improving the learning of the minority class (i.e. neutral). The neutral F-Score has improved from 0.41 (i.e. when using LSTM only), 0.36 (i.e. when using biLSTM only) to 0.44, 0.43 when CNN was combined with LSTM and biLSTM, respectively. This shows evidence that CNN was able to find important features and filter out unimportant ones. Feeding the impor-

tant features to LSTM with bidirectional mechanism has proven to slightly enhance the learning of sequence classification for sentiment analysis on imbalance dataset.

According to our results, we have found that BERT algorithm provides robust capabilities for sequence classification for sentiment and hate speech. It has shown excellent performance in learning binary classification and multi-class classification. In addition, it has been proven effective in dealing with class imbalance as discussed previously in the results.

4.6.2.2.1 English Case Study: Sentiment Behavior Analysis during COVID-19 Pandemic in USA and Canada

Figures 4.10 and 4.11 illustrate the sentiment predictions of our BERT-based sentiment classifier and provide two views of exploratory analysis, temporal and topic-based. Overall during the pandemic, the sentiment tends to be more negative in USA than it is in Canada (as seen in Figure 4.10). The details of COVID-19 data collection for both USA and Canada can be found in Chapter 7.

The temporal analysis of COVID-19 in North America has shown a negative behavior since the very beginnings of the pandemic that is back in December 2019 when the negativity is shown to be higher in Canada than in USA. However, the number of COVID-related tweets during this month was very low (154 tweets) compared to the number of tweets afterwords (4M+ tweets). The negative behavior increased during the months of January and February 2020. Then it decreased over the next three months with Canada exiting the negative zone and entering the positive zone, as depicted in the Figure After May 2020 we again witnessed an increase in the negative behavior for both Canada and USA.

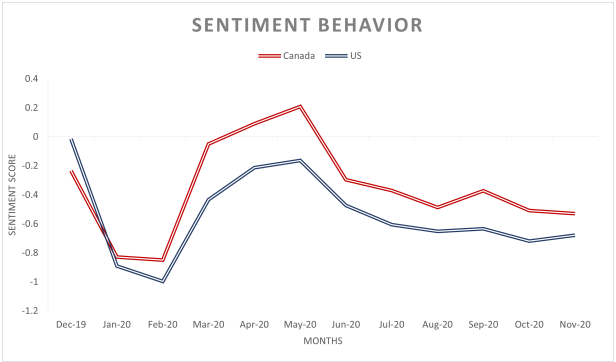


Figure 4.10: **Temporal comparisons of sentiment behavior over twelve months between Canada and USA before and during COVID-19 pandemic.**

To facilitate the understanding of the sentiment behavior over time, we provide a deeper analysis about the topics that people were discussing during the pandemic on OSNs platforms (Figure 4.11). Having the topics at hand, the reasoning of temporal analysis can be achieved. In other words, we can understand the reasons and causes of behavior changes over time. This will help clarifying the story of events. Quarantine topic has shown a high positive behavior in both Canada and USA, for periods 1, 2, and 3, with Canada showing

more positive vibes. This is compatible with the increase in the sentiment positivity in the months of March, April, and May 2020. During these three months, the discussed topics were mostly related to quarantine and staying home. This also explains the low hate behavior that we found in these topics. Our topic model and interpreter confirm this by inferring topics including watching TV, reading books, cooking and baking, as well as providing a set of keywords and phrases (e.g. "fun", "laugh", "dance", "favourite love song listen", "fun food easy food", "stay safe", etc) that people used in the conversations related to these topics. Community support and Health care is anther topic that shows high sentiment positivity. This is not surprising since many of the conversations were related to community and health support including mental health, as shown in the extracted keywords and phrases in Table 7.7 Topic 2 - Period 3 and Table 7.8 - Topic 1 - Period 2. Health-care and front-line workers have provided and received great support through OSNs in USA. This is depicted in the high positive sentiment seen in Figure 4.11(b) - "Public-Health and Support" topic. Some of the supportive tweets were praying to workers and some were from families or friends who have members working in hospital. This is confirmed by the keywords and phrases extracted by our topic model and summarizer. "Pray", "happy", "love family", "hospital worker pray" are examples of the positive vibes that people embraced while interacting with this topic. A high level of negative sentiment is shown in the conversations related to Trump and China, for both Canada and USA during periods 1, 2, and 3. Phrases like "trump hoax", "trump blame china", "president trump beat china" reflect a high negative vibes inferred by our sentiment model. Reporting COVID-19 cases and death toll has also shown negative sentiment throughout the topic discussions with USA showing more negative behavior than Canada did. In addition, topics related to wearing masks have shown a degree of negative behavior in Canada and USA. Moreover, discussing the economy and financial situation during the pandemic appears to have a more negative impact on people in USA than those in Canada.

During period 1 (i.e. Dec 2019 - May 2020) in Canada, our topic model inferred the topic "Public Figures Tested Positive" (Figure 4.11(a)) that discussed mainly famous figures that tested positive. UFC fans, apparently, were not happy about the fighting games being cancelled because of the player Jacare being infected with COVID-19. According to the results, Stephen Miller does not seem to have enough fans in Canada; people showed neither sympathy nor support when he tested positive. Hence, we see the negative reflection in our results. This supports the detected hate behavior found in this topic. 68% of the tweets related to Boris Johnson implied negative sentiment. This again shows an unwelcoming attitude among people in Canada. Topics inferred during period 2 (i.e. May 2020 - Aug 2020) have shown dominant positive sentiments in both Canada and USA (Figures 4.11(c) and 4.11(d)). However, the number of tweets related to COVID-19 (i.e. mentioning covid related keywords) has decreased during June to Aug and the tweets were mostly discussing politics and publich health restrictions. This can be shown in the high negative behavior during months of June till Aug 2020. On the contrary to period 2, period 3 (i.e. Sep - Nov 2020) has generally shown high negative behavior in comparison to the sentiment during period 2 of the pandemic as seen in Figures 4.11(e) and 4.11(f)). Further, topics discussed in USA during this period have mostly shown dominant negative behavior except for "Public Health" and "Trump Supporters" topics showing relatively high positive

sentiments. Canada on the other hand, enjoyed more positive vibes, than USA did during period 3, across its topics with three of which being dominated by positive conversations.

Figure 4.11: **Comparisons of sentiment behavior between Canada and USA over topics inferred during periods 1 & 2 & 3 of COVID-19 pandemic. P1 (Dec 2019 - Apr 2020) is depicted in (a) and (b), P2 (May 2020 - Aug 2020) is dipected in (c) and (d), P3 (Sep 2020 - Nov 2020) is depicted in (e) and (f).**

The main limitation of this work is the non-dynamic topic modeling that does not analyze the evolution of topics over time. Despite this limitation, our proposed topic models have shown good performance results in discovering patterns and inferring topics from the challenging noisy unstructured-formatted OSNs data. Combining TFIDF with NLP techniques and carefully preparing our data and crafting our features have successfully contributed in building topic models that are capable of handling the OSNs data, as reported in our results and analysis. Another limitation is that this work focus on predicting explicit hate and sentiment contents from OSNs messages, however, it was not designed to detect hidden hate or sentiments contained in sarcastic messages. Despite this fact, emojis and emoticons (he iconic features) were considered as an attempt to assist in recognizing hate

and sentiments in this case. An additional limitation of this thesis is data imbalance found in our sentiment dataset (i.e. especially for the neutral class). However, the proposed BERT-based sentiment model provide excellent performance results even on neutral class. We believe that the attention and bidirectional mechanisms adopted by BERT algorithm have minimized the bias towards the majority classes and have shown a very acceptable performance on recognizing the minority class (i.e. neutral). This can be shown in the results reported in this thesis as well as the large-scale analysis.

Chapter 5

Content-Localization based Multi-Lingual-Dialectal Online Social Behavior Modeling

5.1 Introduction

This chapter provides extensive description of the design and implementation of two components of the proposed thesis framework (Figure 3.1) : (1) the "Content-Localization based Translation" component, and (2) the textual sub-component of the content-localization based "Online Social Behavior Analysis" component.

In today's digital and global age, $\approx 52\%$ of internet users access and post information in their own language and dialects. While 75% of world's population do not speak English, French, Arabic and Spanish come as the most spoken languages after English on social media. In order to accurately analyze social behaviors of social users who speak different languages, we need to understand the expressions of the languages they use to share and communicate their opinions and thoughts. Current approaches to solve the multilingual aspect on OSNs is to use language independent resources like textual emojis and emoticons and hashtags [97], use language-dependent data resources [13, 30, 71, 151], or to use translation approach between languages [28, 58, 77]. The first approach could be used as a supplementary tool to assist the analysis of social behavior, but it may miss more important information residing in original texts. The challenge with the translation approach is that it heavily depends on the quality of the parallel translation datasets [211]. Low quality datasets imply low quality models and vice versa [58, 211]. In light of this, while the resource for the English language is fairly sufficient to study different types of social behaviors such as sentiment and hate, similar resources for other languages like Arabic are still immature. The main reason that the resource for Arabic is insufficient is that it has many dialects along with the standard version. Arabs do not use the standard Arabic in their day-to-day communications and conversations. They use the dialectal version of the Arabic language instead. The challenge in this aspect is that each Arabic region speaks a different dialect with different morphology, lexicons, and expressions including idiomatic

expressions. Unfortunately, social users transfer this phenomenon into the digital social media platforms, which in turn has raised an urgent need and great challenges to build suitable datasets and AI models resources for different systems and applications. Decent efforts on building Arabic dialect-based datasets for sentiment, emotion, and hate speech analysis [13, 29, 30, 88, 151] have been found in the research community. However, they reveal a number of limitations: (1) only one type of behavior is considered in building the datasets (e.g. only sentiment or only hate speech) in most of the related researches, (2) only one dialect is considered in building datasets, (3) the size of datasets for training neural networks models is small, (4) the datasets are domain dependent and might not be adapted to other domains, (4) the data collected was restricted to keywords and geographical regions with an assumption that the collected data from a specific geo-region represents the spoken dialect of that region; actually this is not necessarily true as standard Arabic and other dialectal expressions could be unintentionally included, and this may negatively affect the quality of the dataset, and therefore the learning performance for a specific dialect will be affected as well. In this work, we try to overcome the aforementioned issues (i.e., of building new datasets for every language/dialect to solve individual domain OSB problems) by proposing content-localization based translation approach that efficiently allows exploiting existing OSB resources between low-resource and high-resource languages/dialects, and hence minimizing the cost of OSB modeling for individual domains and in specific languages/dialects. The first step towards modeling our translation models is to translate our English domain free multimedia sentiment dataset (DFMSD) into four Arabic dialects (i.e. Gulf, Iraqi, Yemeni, and Levantine/Shami). We do not accept word-to-word translation; instead, we implement a contextual translation of tweets from English to multi-dialect Arabic.

5.2 OSN based Multi-Lingual Multi-Dialectal Arabic Translation Dataset (MLMD)

Our dataset is designed to offer high quality translation models to translate contents on online social networks (OSNs) where people mostly use informal and slang language to express feelings, opinions, and thoughts. The goal of this dataset is to offer a high-quality resource to translate informal and slang English into informal and slang multi-dialect Arabic. Our translation strategy is based on the content localization approach and not word-to-word translation. Content localization translation approach is more precise since it translates contents to be culturally relevant and easy to understand in context. In addition, idiomatic expressions are considered when localizing our English dataset. Moreover, quality control is applied to monitor the translators' work and ensure the quality of the translated datasets. To guarantee the quality of the final translations, we carefully decided on three main aspects: translators, data subsets, and translation strategy.

5.2.1 Translators

We require that all participants are professional translators, native in the Arabic dialects, and native or fluent in English, and familiar with informal and slang English. Also, the translators are required to have a social involvement on social media (i.e. Twitter, Facebook, YouTube or any other platform) using both English and native Arabic dialect. Furthermore, females and males and translators are considered in our translation task to minimize bias in translation; many studies [78, 229] reflect that the way people perceive other's opinions and express theirs are subject to gender differences.

Initially, we wanted to use crowdsourcing platforms to do our translation, but there were not enough native dialect Arabic workers there, and for some dialects there were zero workers. Alternatively, we asked professional translators who are native in Arabic dialects to conduct the translation task. An invitation email was sent to a list of professional translators for the Gulf, Iraqi, Yemeni, and Levantine/Shami dialects. The invitation included a questionnaire asking for their personal and demographic information as well as their social involvement on social media. In order to qualify a translator, we first conducted individual audio calls with potential translators to make sure they are native in their Arabic dialect. Then each candidate undergoes 2-stage qualification test as follows:

1. Qualification Test Stage 1:
   (a) We provide the translator with a list of English test tweets that have been carefully chosen to include slang expressions, social medial language expressions, and idiomatic expressions. We provide the translator with our translation guidelines and ask them to localize the tweets in their native Arabic dialect while preserving the context, tone, and content.
   (b) A corresponding native Arabic dialect speaker will check the translation to approve or decline.
2. Qualification Test Stage 2:
   (a) If the translator passes qualification test stage 1, we provide the first chunk of tweets (i.e. 500 tweets) to be translated, and then we check 50% of the translated tweets randomly.
   (b) A corresponding native Arabic dialect speaker will check the translation to approve or decline. If the translator passes the qualification test stage 2, he/she is qualified as a translator in our translation task.

According to our criteria, 13 out of 39 translators were qualified to participate in our translation task: three translators for Gulf, Yemeni, and Iraqi dialects and four translators for Levantine/Shami dialect. 54% of the participants are females and 46% are males. All the candidates are native in their spoken Arabic dialect, native or fluent in English, familiar with informal and slang English, and have social presence and engagements on social media.

5.2.2 Tweets Subsets

The English tweets to be translated are taken from our DFMSD dataset [1] and ≈ 500 sentences with idiomatic expressions are appended to the tweets list resulting in a total size of 15000 sentences. The data is divided into three subsets, each consisting of 5000 sentences. For each dialect, an individual annotator is asked to translate one subset only (i.e. 5000 sentences) into his/her native Arabic dialect. For the Levantine/Shami dialect, there is one subset of 5000 sentences that has two translators to work on; each has done 2500 sentences. The duration of the translation took around 5 months (starting from June 2021 till October 2021). Our strategy for dividing the data into small subsets and providing the translators with long period of time is to ensure that they feel comfortable while doing this arduous task, hence perform with high level of concentration, carefully understand the sentences, and provide consistent and accurate translations. We have 15000 translated sentences for each dialect, which brings the total of the translated sentences to 60000.

5.2.3 Translation Strategy

For our translation, we adopt the content localization translation approach where translating the texts does not only convey a near-equivalent meaning but also addresses and integrates linguistic, cultural, tone, and contextual components of the texts. Same words might convey different meaning in different dialects; for example, the word “chips” refers to fried thin potato chips in North American English while it refers to “fries” in UK English. The same applies for different dialects in Arabic language. An example of this, the word "صاحبي" in Gulf dialect refers to a friend while in Lebanese dialect it refers to boyfriend.

Since the dataset has been collected from social media (i.e. Twitter in this work) and the tweets are mostly expressed in a day-to-day spoken language, the translation from English to multi dialectal Arabic will be customized to OSN cultural language. Thus, we take into consideration a number of additional criteria (i.e. in addition to the content localization based translation approach):

1. (1) Consideration of OSN cultural language and expressions: iconic emotion (e.g. emoticons, emojis), slang abbreviations, and hashtag words should be kept in the translated texts while preserving their occurrence order and their context. Hashtag words are translated into the corresponding Arabic dialect.

2. (2) Consideration of Informal Language: informal language is used in casual settings like in daily conversations. It includes slang words like "lol", abbreviations like "idk", and usually shorter sentences like "imo ur walking fast" or "OMG, I just got a new job!".

3. (3) Consideration of idiomatic expressions: an idiomatic expression should not be word-to-word translated. Instead, it should be translated to convey the context or its equivalent idiomatic expression in the corresponding dialect while preserving the context of the original text.

4. (4) Consideration of language code borrowing: code borrowing refers to using one primary language but mixing in words from another language to fit the primary language. For example, the word "lol" is used and written using the Arabic alphabet as "لول" ; similarly, the word "pistachio" is used and written using the Arabic alphabet as " بيستاشيو"

In this work, four main Arabic dialects are studied: Gulf, Yemeni, Iraqi, and Levantine/Shami. The qualified Translators are provided with a subset of English sentences and translation guidelines that cover the above criteria. In addition, they are advised to convert mainly proper nouns into Arabic letters where applicable – for example names of people ("John" to "جون"), and names of places ("Lebanon" to "لبنان"). The translators are also advised to pay attention to the spelling as any misspelling would harm the quality of the translation. Upon the completion of the translation task, the translators are asked to do a round of proofreading before they submit the final translations. Note that the translation is done for each dialect individually with the corresponding dialect translators.

5.3 Content-Localization based Neural Machine Translation (NMT) Modeling

5.3.1 Preprocessing

- Removing extra whitespaces
- Removing encoding symbols
- Removing user mention
- Removing URLs
- Converting text to lower case: it is applied to English words existing in the data.
- Removing special characters and numbers
- Removing tashkeel and harakat: tashkeel or harakat refer to all the diacritics placed over or below letters. For example, "بُسْتان" becomes بستان",
- Normalizing Hamza: Normalizing Hamza forms into one form. For example: "أهلا" becomes "ءهلا".

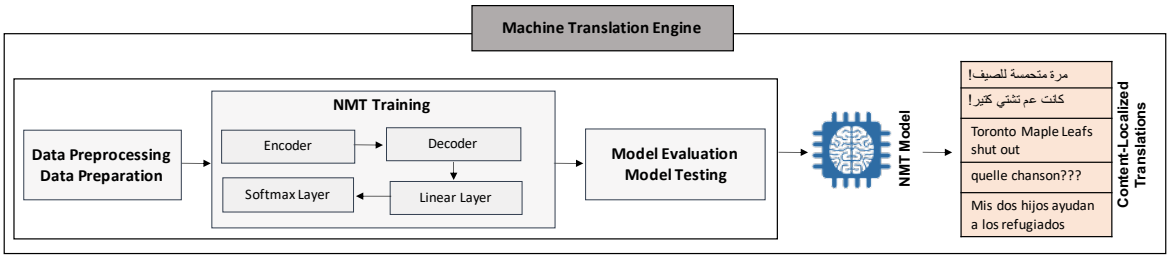


Figure 5.1: **Neural machine translation framework.**

5.3.2 Sequence-to-Sequence Transformers Model

We use the sequence-to-sequence Transformers architecture depicted in Figure 5.2 [130]. It consists of 12 layers of encoder and 12 layers of decoders with model dimension of 1024 on 16 heads. On top of both encoder and decoder, there is an additional normalization layer that was found to stabilize the training. We use pretrained weights from mBART [130,209] model that was pre-trained on 25 languages. It has been proven that transfer learning offers a rich set of benefits including improving the efficiency of model training and saving of time and resources since building a high-performance model from a scratch requires a large amount of data, time, resources, and efforts. Therefore, we use the fine-tuning learning approach [17,220] to train Transformers machine translation models as an attempt to solve the limitations [19] of small datasets and domain adaptation. For our Arabic dialects models, we use our proposed multi-dialectal datasets to finetune the mBART [130] pretrained model. The model learns parallel translation from informal and slang English to informal and dialectal Arabic. We have four models corresponding to four dialects: Iraqi, Yemeni,Gulf,and Levantine/Shami. Figure 5.1 illustrates the general framework followed in training our neural machine translation models.

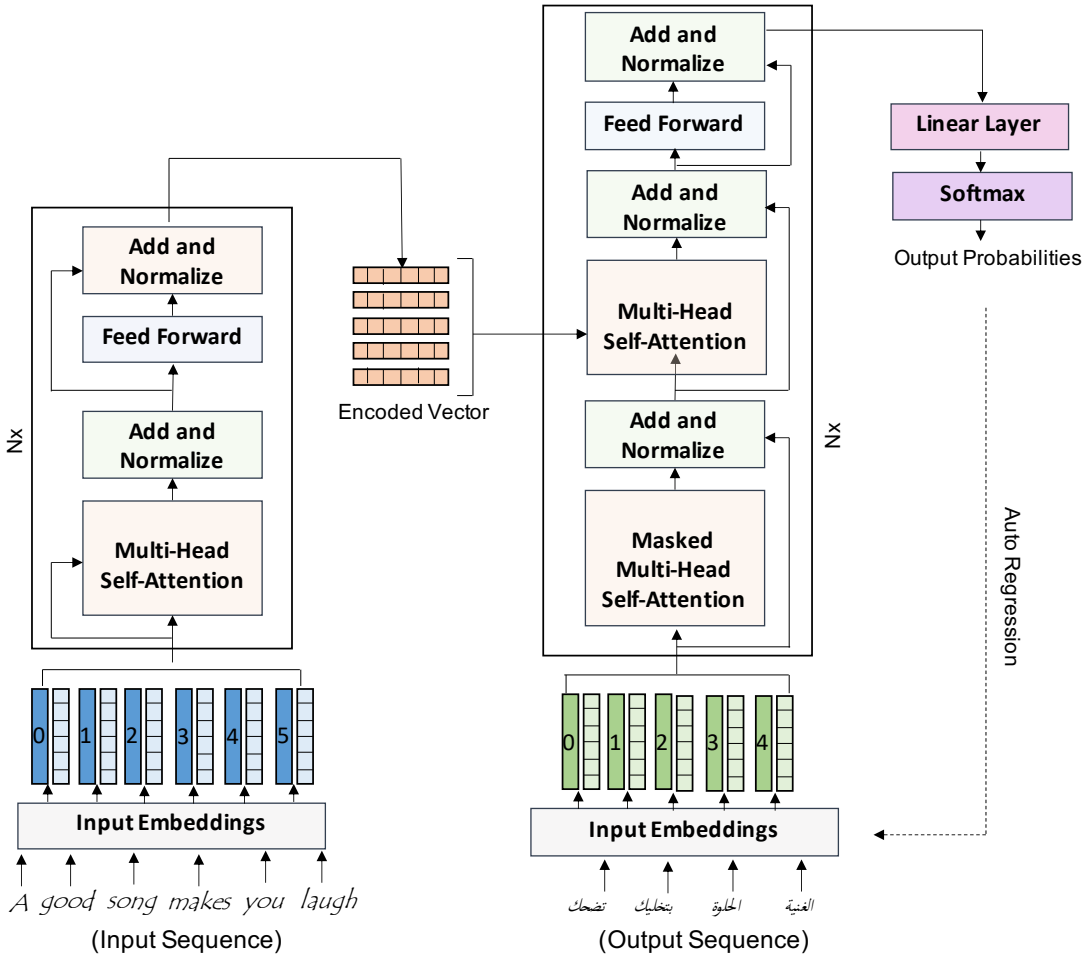


Figure 5.2: **Sequence-to-sequence Transformers architecture [216] used in training our multi-lingual-dialectal NMT models.**

5.3.3 Experimental Results & Analysis

The purpose of this section is to assess the performance of our proposed NMT models. We present the experimental design, evaluation protocol, and then we report results and discuss findings.

5.3.3.1 Experiment Design and Evaluation Metrics

For evaluation purposes, we design four experiment scenarios:

- We train sequence-to-sequence (seq2seq) with attention models using our proposed dataset; one model for each dialect of four. We then compare the performance of seq2seq models with our proposed Transformers based models.
- We conduct a two-way cross-dialect evaluation of our proposed NMT models (i.e. English-Arabic) on the four Arabic dialects (Levantine, Gulf, Iraqi, and Yemeni) using the test sets of the four models. Each model was tested on a different test set that was chosen randomly. The purpose of this experiment is to assess the quality of our proposed MLMD dataset and to investigate the differences of dialects within a language (i.e. Arabic in this research) in order to examine whether the translation quality of messages within the same language is affected.
- We conduct transitive translation evaluation from non-English languages to Arabic dialects. We use transitive property as following:

  $$\text{if } A = B \text{ and } B = C, \text{ then we know } A = C$$

  We adopt the transitive property to fit our case as following:

  let's assume A = English sentences, B = Arabic translation of A, C = Spanish translation of A, D = French translations of A, then

  $$\text{if } B -> A \text{ and } C -> A, \text{ then } B -> C$$

  Similarly,

  $$\text{if } B -> A \text{ and } D -> A, \text{ then } B -> D$$

  We manually translate a subset of English sentences (i.e. from our MLMD dataset) to Spanish and French. Then, we train NMT models from/to Spanish and French to/from Arabic dialects [20].
- We conduct an evaluation comparison between our proposed NMT models against GPT3.5 [44] and GPT4 [166] models as an attempt to assess the translation performance of our NMT model.

For the evaluation metrics, we use BLEU and ROUGE metrics. BLEU metric is recommended for the translation tasks as it conducts robust assessment over the quality of translation fairly quickly. ROUGE complements BLEU in terms of evaluation where it focuses on recall; while BLEU is precision oriented. Therefore, we use F-Score of BLEU and ROUGE.

The data was split with ratio of 90% for training and 10% for testing. The default settings of mBART were adopted during model training.

5.3.4 Results and Analysis

5.3.4.1 Performance of the Proposed NMT Models

Table 5.1 shows that Transformers-based models outperform seq-to-seq with attention models that were trained without transfer learning approach. Transformers models that were trained through finetuning a pretrained model (i.e. mBART in this work) yield by far better translation with an improvement of $\approx$ 20 points for Yemeni, 37 points for Iraqi, $\approx$ 35 points for Gulf, and $\approx$ 33 points for Levantine/Shami.

The nature of Transformers that use multi-head self-attention allows the models to learn better word contextualization, and hence yielded better results in terms of F-Score of BLEU and ROUGE. The self-attention enables contextualizing every word in various positions with respect to the whole sequence. This solves the homonym problem where similar words might have different meanings in different contexts. This problem was shown in the performance of seq-to-seq with attention models (Table 5.1) where attention is used to connect recurrent states of encoder with the recurrent states of decoder. In addition, learning a model from scratch requires a huge size of data so that the model can converge properly. However, transfer learning technique has solved the problem related to limited and/or small data resources. As seen in Table 5.1, leveraging the transfer learning in learning our NMT models has tremendously improved the models' performance.

Table 5.1: **Comparison results between sequence-to-sequence with attention and sequence-to-sequence Transformers -in terms of F-Score(BLEU, ROUGE)- using our proposed multi-lingual-dialect dataset. The results represent the English to Arabic dialects models.**

	Yemeni	Iraqi	Gulf	Levantine/Shami
seq-to-seq with attention	11.3	2	13.1	2
seq-to-seq Transformers	31.4	39.8	34.9	34.6

5.3.4.2 Cross-Dialect Evaluation of the Proposed NMT Models

Tables 5.2 and 5.3 illustrate the results of eight models translated from/to English to/from four Arabic dialects. English to Arabic-Levantine model (Table 5.2) performs the best on

Table 5.2: **Performance results of four English to Arabic-dialect models -in terms of F-Score(BLEU, ROUGE)- using our proposed multi-lingual-dialectal dataset. The four models represent machine translators from Arabic-Levantine, Arabic-Gulf, Arabic-Iraqi, and Arabic-Yemeni, to English.**

		Test Set	En -> Ar-Levantine	En -> Ar-Gulf	En -> Ar-Iraqi	En -> Ar-Yemeni
Model	En -> Ar-Levantine	Test Set-1a	**34.6**	29.6	21.1	21
	En -> Ar-Gulf	Test Set-2a	28.4	**34.9**	22.6	27.3
	En -> Ar-Iraqi	Test Set-3a	19.9	22.8	**39.8**	20.8
	En -> Ar-Yemeni	Test Set-4a	19.1	28.1	20.1	**31.4**

Table 5.3: **Performance results of four Arabic dialects to English models -in terms of F-Score(BLEU, ROUGE)- using our proposed multi-lingual-dialectal dataset. The four models represents machine translators from English to Arabic-Levantine, Arabic-Gulf, Arabic-Iraqi, and Arabic-Yemeni.**

		Test Set	Ar-Levantine -> En	Ar-Gulf -> En	Ar-Iraqi -> En	Ar-Yemeni -> En
Model	Ar-Levantine -> En	Test Set-1b	**51.3**	45.1	40.1	40.3
	Ar-Gulf -> En	Test Set-2b	43	**48.5**	40.3	42.3
	Ar-Iraqi -> En	Test Set-3b	38.2	39.3	**50**	37.1
	Ar-Yemeni -> En	Test Set-4b	39.7	40.5	37.2	**43.2**

the Levantine test set with F-Score of 34.6 that is better by $\approx$ 5-14 points when compared to its performance on the other dialects. The same observations are seen in the English to Arabic-Gulf, English to Arabic-Iraqi, and English to Arabic-Yemeni with F-Scores of $\approx$ 35, 40, and 31.4, respectively. The results of Arabic dialects to English are shown to be consistent with the results of English to Arabic dialects where dialectal models perform the best on its native test set.

It can be observed -from the results in both tables (5.2 and 5.3)- that Levantine models are closer to Gulf models than Iraqi and Yemeni models are, whereas Gulf models are closer to Levantine and Yemeni models than to Iraqi models. Iraqi models are shown to be closer to Gulf than to Levantine and Yemeni, while Yemeni models are closer to Gulf than to Levantine and Iraqi. Actually, these findings reflect the real-life fact that Levantine and Gulf people can understand each other's dialects better than they understand Iraqi's and Yemeni's. This is due to that fact that Levantine and Gulf dialects use similar words but different pronunciations more than in the case with the Iraqi dialect. On the other hand, Iraqi dialect shares more words with Gulf dialect than Levantine and Yemeni dialects do. Similarly, Yemeni people can understand the Gulf dialect better than the Iraqi's and Levantine's.

Our results show that there are differences in expressions between dialects within the same language (i.e. Arabic in this thesis) and that is shown in models performing the best on their native data. A model of a particular dialect is shown to be capable of preserving the context of messages when translating the messages to/from its native dialect, while it might miss transferring the context to the other non-native dialects (as shown in Tables 5.2 and 5.3). This finding highlights the importance of understanding the information in its native language and dialect while preserving the context at the same time. This is crucial

for decision makings in applications where the precision of understating information impact the analysis of the problems in hands (i.e. online social behavior analysis in this thesis).

5.3.4.3 Performance Comparison between the Proposed NMT Models and Large Language Models (LLMs)

Tables 5.4 and 5.5 show performance comparisons of our English-dialectal Arabic (i.e., Levantine and Gulf dialects) NMT models against the translation performance of GPT3.5 [44] and GPT4 [166] models. The results show that our models outperform the GPT3.5 and GPT4 models in translating from/to English to/from informal dialectal Arabic using tweets. Our results have also revealed that GPT3.5 and GPT4 models are not capable of translating to informal dialectal Arabic as evidenced by their poor performance in term of F-Score(BLEU ROUGE) metric (Table 5.4). In addition, our NMT models are shown to be better at understanding and translating local Arabic dialects into English compared to GPT3.5 and GPT4 models. This is evident from the reported results, which demonstrate the superior performance of our proposed NMT models (Table 5.5). The observed performance behavior of GPT models might be due to the facts that these LLMs have been dominantly trained on English data. The limited exposure of LLMs to dialectal Arabic data might hinder their ability to capture the unique expressions, grammar, and cultural references prevalent in these dialects, leading to unnatural and inaccurate translations as reported in Tables 5.4 and 5.5. It is important to mention that our NMT models, GPT3.5, and GPT4 models were evaluated on the same test set. This ensures a fair comparison and therefore allows for a more accurate performance assessment of our proposed NMT models.

Table 5.4: **Comparison of the translation performance between our proposed NMT models against GPT models (GPT3.5 and GPT4) using informal social media messages (i.e., tweets). The tweet translation is from English to four Arabic dialects.**

Ar-Dialect	**Model**	**F-Score(BLEU,ROUGE)**
Levantine/Shami	GPT4	0.072620469
	GPT 3.5	0.061804148
	Our model	**34.6**
Gulf	GPT4	0.067356286
	GPT 3.5	0.058850802
	Our model	**34.9**
Iraqi	GPT4	0.046998601
	GPT 3.5	0.043390178
	Our model	**39.8**
Yemeni	GPT4	0.057397155
	GPT 3.5	0.050376419
	Our model	**31.4**

Table 5.5: **Comparison of the translation performance between our proposed NMT models against GPT models (GPT3.5 and GPT4) using informal social media messages (i.e., tweets). The tweet translation is from four Arabic dialects into English.**

Ar-Dialect	Model	F-Score(BLEU,ROUGE)
Levantine/Shami	GPT4	0.2737502
	GPT 3.5	0.218327815
	Our model	**51.3**
Gulf	GPT4	0.242276302
	GPT 3.5	0.195508991
	Our model	**48.5**
Iraqi	GPT4	0.21725176
	GPT 3.5	0.170278056
	Our model	**50**
Yemeni	GPT4	0.206334108
	GPT 3.5	0.166573635
	Our model	**43.2**

Figure 5.3 illustrates a list of example translations generated by our proposed models and Google Translate [103], from/to English to/from Arabic four dialects. Please note that all the sample translations were examined and approved by bilingual people with native to near-native bilingual language proficiency. In Figure 5.3, we demonstrate that our proposed models have successfully met the four criteria they were designed to meet (as discussed earlier in this chapter), as follows:

- **Content localization:**

  Sentences 8 and 12 show how the idiomatic expression "the exam was a piece of cake" is localized differently to the Levant and Gulf dialects. Both dialects use different terms to express the same sentence in English. Levantine dialect uses term "هين" to describe "easiness" while Gulf uses term "سهل" to describe "easiness". Also, the word "exam" means "فحص" in Levantine dialect while it is "اختبار" in Gulf dialect. The word "فحص" has another meaning in Gulf dialect and it means "to check, to examine, or to inspect" and it is never used to describe "exam" as it is the case in Levantine dialect. Sentences 5, 18, and 23 are other examples of content localization between dialects. Levantine, Gulf, and Iraqi describe "raining heavily" differently; they use different words to describe "raining" and different exaggeration terms. For

"raining", Levantine uses "تشتي", Gulf uses "تمطر", and Iraqi uses "تزخ". To express exaggeration, Levantine uses "كتير" whereas Iraqi uses "كومة". Moreover, month "May" is expressed as "أيار" in Levantine while it is "مايو" in Gulf and Yemeni, as seen in sentences 9, 16, and 28. Sentence 19 also shows how the model localizes the sentence; it knows that the corresponding term "cats" is "بسس" and "now" is expressed as "الحين" in Gulf dialect. Sentence 7 illustrates that the model successfully transferred the context of the sentence from Levantinian dialect to English. The real translation of sentence 5 "ديما عبارة عن شخص هبيل مع دماغ مستطيل" is ""Dima is an idiot person with an empty brain". The literal meaning of "دماغ مستطيل" is "rectangular brain" while it its real meaning is "empty brain, ignorant or not very intelligent". Our model did no do a literal translation; instead it translated "دماغ مستطيل" as "retarded brain" which is correct.

- **Consideration of OSN cultural language and expressions:**

  Sentences 2, 3, 26 show that our models are capable of handling OSN culture language like hashtags, and slang abbreviations. sentence 2 "kinda finished my work" was correctly localized to its Levantine corresponding translation "نوعا ما خلصت شغلي". The slang word "kinda" was correctly translated as "نوعا ما". "kinda" in sentence 3 was also correctly translated to شوي" which means "a little bit" to describe "kinda bored" as "a little bit bored" with the expression "شوي زهقان. رح ءلعب بي ءس ه". In addition, the model could recognize the slang abbreviation "Ima" and localized it to "رح" which means "I will" in Levantine dialect. Sentence 20 "sorry idk" contains an abbreviation for "I don't know" in which our model detected and correctly translated it into its Gulf expression "ماعرفه". In sentence 26, hashtag "#KeepFighting-Michael" is three words ("Keep", "Fighting", and "Michael") written in a one word. Our model could recognize the three words and correctly translated them into its corresponding Iraqi dialect and connected them by a "_" as "#ابقة_الي_تقاتل_مايكل". Similarly, in sentence 33 where hashtag "#DavidDraiman" and "#TheLight" are composed hashtags. Our model recognized them, but here it preserved their English names and only converted the English letters into Arabic alphabets and connected them by "_" as "#ديفيد_دريمان" and "#ذا_لايت"

- **Consideration of Informal Language:**

  Sentence 10 "let's try booking for a flight, fingers crossed it works out" is localized to a Gulf translation as "خلونا نحاول نحجز رحلة، انشالله رح تكون كويسة". The context of informal phrases "fingers crossed" and "works out" were correctly transferred to the translated sentence as "انشالله" which means "God willing" and "تكون كويسة" which means "to be good". Similarly, in sentence 1 where expressions "what the hell!" and "weirdo" were translated to its Levantine expressions "شو هالشغلة!" and

"غريبة". The Levantine expression "شو هالشغلة!" indicates an exclamation of anger, annoyance, or surprise and its literal meaning is "what is this thing". The expression in sentence 11 "he's screwed!" is an informal way of saying that a person is in trouble. Our model translated it into "جاب العيد!" which is a common local Gulf way of expressing that someone is in a bad situation when things gone wrong. The literal meaning of "جاب العيد!" is "he brought Eid celebration" which has nothing to do with its real context. Note that Eid is religious celebration for Muslims which is celebrated twice a year: the Fitr and the Adha. The term "utterly spectacular" was localized to Gulf expression "يجنن مرة" which is, again, a common local way in Gulf to express how beautiful things are; the literal translation of "يجنن مرة" is "very crazy" which has a different meaning in English and is far from its real context. "مرة" is an exaggeration term corresponding to "very" in which it is locally used in Gulf and its literal meaning is "once". In sentence 6, "I'm craving" was localized to the Levantine expression of "طالع ع بالي" that has a literal meaning of "up in my mind". Sentence 22 shows an example of an Iraqi way of saying "dog" as "جلب" as Iraqis pronounce alphabet "ك" as "ج" whereas other dialects do not. The original term for "dog" across majority of Arabic dialects is "كلب" especially in Levantine and Yemeni dialects; they never pronounce "ك" as "ج". Fortunately, our models have properly learnt dialects and our Iraqi model could successfully localize the translation of "الجلب هذا جدا جميل" into "this dog is so beautiful" while the Levantine model was not able to detect the whole context and translated the sentence into "this ice is so beautiful"; it translated term "جلب" as "ice" instead of "dog".

- **Consideration of idiomatic expressions:**

  Our models show effective handling of translating idiomatic expression across the four Arabic dialects. The English idiom "raining cats and dogs" was properly localized into Levantine, Gulf, and Iraqi dialects (sentences 5, 18, and 23). Although the literal meaning of this idiom refers to the animals cats and dogs, our models have effectively learnt its actual context and not its literal meaning. Interestingly, our model could translate the Arabic-Gulf expression ("بتمطر غزير") of "raining heavily" into an English idiomatic expression as "it's raining like cat and dog". Another example is the idiomatic expression "im over the moon" that actually means "extremely happy" but its literal meaning " is to be literally over the moon". Our model was able to learn the idiom's context and successfully transferred the actual context into a Yemeni translation as "قدنا متحمس"; "متحمس" in Yemeni dialect means "excited". "no-brainer" is also an English idiomatic expression with a literal meaning of "no brain", but actually it means "very easy". Our model learnt the context meaning of "no-brainer" and translated it into "قرار سهل", which means "an easy decision". In sentence 32, the English idiom "no pain no gain" was successfully localized into the Yemeni dialect as "مابش ربح بدون تعب" where words "مابش، بدون" are used

locally by Yemenis to say "no", "ربح" which means "winning" to say "gain", and "تعب" means "tiredness" to say "pain".

- **Consideration of language code borrowing:**

  OSN slang expression "looool" (in sentence 4) was correctly detected as a laughter or funny expression and was translated as "لوول" which is the same English word (i.e. pronounced in Arabic similar to its English pronunciation but expressed in Arabic alphabets. Similarly, in sentence 3 where the abbreviation "ps5" (i.e. PlayStation 5 video game device) was translated to its corresponding Arabic alphabets ("بي ءس ه") while preserving its pronunciation in English. Sentence 33 follows the same translation pattern where the word "#TheLight" translated into its corresponding Arabic alphabets "#ذا_لايت". The name "gaku shibasaki" was also converted to its corresponding Arabic alphabets as "جاكو شيباساكي".

Figure 5.3 illustrates how well our proposed NMT models perform against Google Translate at translating informal social media messages from/to English to/from four Arabic dialects (i.e., Levantine/Shami, Gulf, Iraqi, and Yemeni). The table shows that our NMT models were able to localize the tweets between English and dialectal Arabic in term of idiomatic expressions, informal local expressions, social media-cultural expression, and language code borrowing. In contrary to our NMT model, Google Translate did not produce dialectal Arabic translations; instead, it only translated English tweets into Modern Standard Arabic (MSA). Google Translate has also shown a weaker performance at understanding idiomatic expressions and hence producing word-to-word translation as seen in the examples "it was raining cats and dogs" meaning "it was raining heavily", "the exam was a piece of cake!" meaning "the exam was easy!", and "ديما عبارة عن شخص هبيل مع دماغ مستطيل" meaning "Dima is a stupid person with dump brain", which were word-to-word translated by Google Translate; unlike our NMT models that have successfully produced contextual translations of these examples. Informal local expressions and social-media cultural expressions were shown to be mistranslated by Google Translate; "he's screwed" was translated into Arabic as "he is drunk", "sorry idk him" was translated into Arabic as "sorry I know him", and "بنت ترمب يا عيال" meaning "trump's daughter, guys" was translated from Gulf-Arabic into "Trump's daughter, kids". Unlike our NMT models that were able to translate "الجلب هذا جدا جميل" from Iraqi-Arabic into English as "the dog is so beautiful", Google Translate was not able to recognize the Iraqi dialect that uses word "الجلب" as equivalent to the English word "dog". While Google Translate correctly translated "بتمطر غزير" into "it is raining heavily", our NMT models took the translation further and almost localized "بتمطر غزير" into the English idiomatic expression as "it's raining like cat and dog". Levant people use word "ايار" to describe the month of "May" while Gulf people use word "مايو". Our NMT models were able to localize the translation of month "May" into its local equivalent for both Levant and Gulf dialects while Google Translate translated "May" as "مايو" for both Arabic dialects.

	Original Sentence	Google Translator	Our Models' Translation	Dialect
1	what the hell! she is a weirdo	بحق الجحيم! إنها غريبة الأطوار	شو هالشغلة! هيي غريبة	Levantine
2	kinda finished my work	أنهيت عملي نوعًا ما	نوعا ما خلصت شغلي	Levantine
3	kinda bored. Ima play ps5	أشعر بنوع من الملل. انا العب ps5	شوي زهقان. رح علعب بي عس 5	Levantine
4	Looool it's so funny	لوول إنه مضحك جدا	لوول كتير بيسلي	Levantine
5	it was raining cats and dogs	كانت تمطر قططا وكلابا	كانت عم تشتي كتير!	Levantine
6	Im craving a tuna sandwiche!	أنا أشتهي شطيرة التونة!	كتير طالع ع بالي ساندويش تونا!	Levantine
7	ديما عبارة عن شخص هبيل مع دماغ مستطيل	Dima is a stupid person with an oblong brain	always a dumb person with a retarded brain	Levantine
8	The exam was a piece of cake	كان الامتحان قطعة من الكعكة	الفحص كان شي كتير هين	Levantine
9	Shes graduating in May	انها تخرج في مايو	حتتخرج بأيار	Levantine
10	lets try booking for a flight, fingers crossed it works out	لنحاول حجز رحلة طيران، لقد نجح الأمر	خلونا نحاول نحجز رحلة، انشالله رح تكون كويسة	Gulf
11	he's screwed!	لقد ثمل!	جاب العيد!	Gulf
12	the exam was a piece of cake	كان الامتحان قطعة من الكعكة	الاختبار كان شي سهل	Gulf
13	Im feeling down today	أشعر بالإحباط اليوم	زعلانة اليوم	Gulf
14	بنت ترامب يا عيال	Trump's daughter, kids	trump girl y'all	Gulf
15	بتمطر غزير	It's raining heavily	it's raining like cat and dog	Gulf
16	She's graduating with a bachelor's degree in May	ستتخرج بدرجة البكالوريوس في شهر مايو	حتتخرج بدرجة بكالوريوس في مايو	Gulf
17	california dreamin' of with training on the pacific coast - utterly spectacular! #teamredontour	تحلم كاليفورنيا بالتدريب على ساحل المحيط الهادئ - مذهل تمامًا! #teamredontour	جالسة أحلم بالتدريب على ساحل المحيط الهادي- شي يجنن مرة! #جولة_الفريق_ريدون	Gulf
18	it was raining cats and dogs	كانت تمطر قططا وكلابا	تمطر بغزارة	Gulf
19	mission cats in a box complete. now for a trip to the vets. they are not happy!	مهمة القطط في صندوق كاملة. الآن لرحلة إلى الأطباء البيطريين. هم ليسوا سعداء!	مهمة البسس في الصندوق خلصت. الحين لرحلة للزبط البيطري. ما فرحوا!	Gulf
20	sorry idk him	آسف أعرفه	اسف ماعرفه	Gulf
21	it was raining cats and dogs	كانت تمطر قططا وكلابا	تمطر بغزارة	Gulf
22	الجلب هذا جدا جميل	This fetch is very beautiful	this dog is so beautiful	Iraqi
23	It's been raining cats and dogs all day. I'm afraid the roof is going to leak.	لقد كانت السماء تمطر القطط والكلاب طوال اليوم. أخشى أن السقف سوف يتسرب.	دا تزخ كومة كل اليوم. اني خايف انو السكف راح يتسرب.	Iraqi
24	Now people are watching the replay. I didn't even do anything! #Scared of #Periscope. 👀 Haha!	الآن الناس يشاهدون الإعادة. لم أفعل أي شيء حتى! #الخوف من #البيريسكوب. 👀 هاها!	هسه الناس دا يشوفون ري بلي. اني مسويت اي شي! #قلقيت من #بيريسكويب. هههههه!	Iraqi
25	no-brainer	بدون تفكير	قرار سهل	Iraqi
26	Currently Michael Schumacher's weight is arround 45kg!!!! #Unbelievable #KeepFightingMichael	يبلغ وزن مايكل شوماخر حاليًا حوالي 45 كجم !!!! #لا يصدق #واصل القتال مايكل	حاليا وزن مايكل ماكشور حوالي 45 كغ!!! #مدا_اصدك #ابقة_الي_تقاتل_مايكل	Iraqi
27	gaku shibasaki now playing fantastic levels ... will give best suppoto japan national team ... I wish all best	يلعب جاكو شيباساكي الآن مستويات رائعة ... سيقدم أفضل دعم لمنتخب اليابان الوطني ... أتمنى كل التوفيق	جاكو شيباساكي يلعب الان بمستويات رهيبة... شتجيب ءفضل دعم للءعضاء اليابانية... ادمنى كل الخير	Yemeni
28	She's graduating with a bachelor's degree in May	ستتخرج بدرجة البكالوريوس في شهر مايو	بتتخرج بدرجة البكالوس في مايو	Yemeni
29	I am feeling down today!	أشعر بالإحباط اليوم!	انا حاسس بشعور خايس اليوم!	Yemeni
30	I got the job im over the moon	لقد حصلت على الوظيفة وأنا فوق القمر	لقيت وظيفة قدنا متحمس	Yemeni
31	he's screwed our day!	لقد ثمل يومنا!	ضيع يومنا!	Yemeni
32	No pain no gain	لا ألم, لا ربح	مابش ربح بدون تعب	Yemeni
33	Dont let hope become a memory #DavidDraiman #Disturbed #TheLight	لا تدع الأمل يصبح ذكرى #ديفيد درايمان #مضطرب #النور	لاتخلي الامل تصير ذكرة #ديفيد_دريمان #مستاء #ذا_لايت	Yemeni

Figure 5.3: **A sample of generated translations by our proposed NMT models and Google Translate [103]. The translations are from/to English to/from four Arabic dialects: Levantine, Gulf, Iraqi, and Yemeni.**

5.3.4.4 Transitive-Translation to Arabic-Dialects Evaluation

Table 5.6 presents the results of French and Spanish to/from Arabic dialects NMT models [20]. Spanish to Arabic-Gulf NMT model is shown to have the highest learning performance (F-Score of 37 points) followed by Spanish to Arabic-Gulf (F-scoe of 36 points) and then French to Arabic-Levantine (F-Score of 35 points). The F-Score of the models are $\geq$ 35 points and that is comparable to the ones of the English to Arabic dialects models. This result confirms the quality of our proposed MLMD dataset and that is shown in the models ability to learn the translation of the Arabic dialects even with the transitive translation approach (i.e. English sentences were translated into French and Spanish, and then the Arabic translations of English sentences were associated with the French and Spanish translations).

Table 5.6: **Performance results of three transitive-translation based models -in terms of F-Score(BLEU, ROUGE)- using our proposed multi-lingual-dialectal dataset. The three models represent Spanish to Arabic-Levantine, Spanish to Arabic-Gulf, and French to Arabic-Levantine.**

		F-Score(BLEU, ROUGE)
Model	Es -> Ar-Levantine	37.1
	Es -> Ar-Gulf	36.2
	Fr-> Ar-Levantine	35.2

Please note that a subset of our MLMD English sentences were manually translated into French and Spanish languages. That is due to the time constraint and difficulties we faced in finding professional native or near native bi-lingual translators that are committed to work on a large dataset. The subset of the sentences was randomly chosen but do not contain idiomatic expressions. To the best of our knowledge, this is the first work which proposes an NMT model that translates OSN messages from/to French and Spanish to/from Arabic dialects. Levantine and Gulf dialects were chosen with French and Spanish language for the purpose of evaluating our proof-of-concept- case studies conducted in Chapter 4.

Figure 5.4 illustrates a sample of translations (i.e. generated by our proposed models) from/to French and Spanish to/from Arabic dialects (i.e. Levantine and Gulf in this case study). It is important to note that all the sample translations were examined and approved by bilingual users with native to near-native bilingual language proficiency. The participant users approved that the contexts of the original messages were properly transferred -along with the corresponding dialect expression- into our generated translations. "Oh mon Dieu!" is an informal expression to say "oh my god or for god sake!" which was correctly translated into the Levantine expression "!يارب" that indicates an exclamation of anger, annoyance, or surprise. "bizarre" is a common French word to describe someone or something as "weird or strange". The expression "الفحص" is a very local way of saying "exam" in Levantine dialect, which could mean "check or examine" in other dialects like Gulf and Yemeni. Our Levantine

model could successfully recognize it as "exam" and generate the sentence translation as "L'examen était très facile". "L'examen" in French refers to "exam" in English. Similarly, with expression "تشتي" which means "raining" in the informal Levantine dialect; our model correctly translated it into French as "il pleuve tellement". "pleuve" is the French word for "raining".

	Original Sentence	Our Models' Translation	Language	Dialect
1	Oh mon Dieu! elle est bizarre	يا ربي! هيي غريبة	French	Levantine
2	J'ai terminer mon travail.	خلصت شغلي		
3	Je voudrais un sandwich au fromage.	بدي سندويش جبنة!		
4	Dima est un personne stupide avec un cerveau retardé.	ديما شخص غبي مع دماغ متءخر		
5	الفحص كان كتير هين	L'examen était très facile		
6	كانت عم تشتي كتير	il pleuve tellement		
7	Los expertos nos informan constantemente de que pasamos demasiado tiempo mirando la televisión y nos advierten que eso no es bueno para la salud	الخبراء بيخبرونا دايما ءنو منقضي وقت طويل عم نحضر التلفزيون ويحذرونا ءنو هيدا مش منيح للصحة	Spanish	Levantine
8	dos de los amigos de eli trabajaron de forma independiente durante toda la semana ilustrando y escribiendo este libro de dragones, lo hicieron durante su tiempo libre	اشتغلت تنين من رفقاتي لـ ءيلي بشكل مستقل كل الءسبوع عم يصورو و يكتبو هالكتب التنينات عملو هالشي بوقت فراغهم		
9	¡Respira profundo! No te preocupes. Estoy aquí para ayudarte.	استرخي عميق! ما تقلق. ءنا هون لءساعدك.		
10	Tan emocionado por el verano.	كتير متحمسة للصيف		
11	شي كتير عظيم أنو كون موجود بهالحدث الحلو وشوف شو عم يصير	¡es tan bueno estar presente en este evento y ver lo que está sucediendo!		
12	بتسم كل يوم و أنا متأكد انك رح تكون مبسوط. اجبور حالك! #مبسوط #قلق #يأس	te sonríes todos los días y estoy seguro de que estarás feliz. ¡adrígate! #feliz #ansiedad #desesperanza		
13	Los expertos nos informan constantemente de que pasamos demasiado tiempo mirando la televisión y nos advierten que eso no es bueno para la salud	'الخبراء دايما يخبرونا اننا نقضي وقت كثير نطالع التلفزيون و يحذرونا انه هذا مو كويس للصحة'	Spanish	Gulf
14	dos de los amigos de eli trabajaron de forma independiente durante toda la semana ilustrando y escribiendo este libro de dragones, lo hicieron durante su tiempo libre	اثنين من ءصحابه اشتغلو بشكل مستقل طول الءسبوع ينرسمون و يكتبون كتاب التنين ذا سووه في وقت فراغهم		
15	Vi esta película durante el fin de semana - ¡estuvo genial! #willsmith estuvo increíble.	شفت هذا الفيلم في عطلة نهاية الءسبوع- كان راءع! #ويل_سميث كان راءع. #مضحك #صور_تضحك		
16	Mi teléfono duró toda la noche. #4s #impresionado	جوالي جلس طول الليل. #4_س #منبهرة		
17	في عصر المعلومات الجهل خيار #شبكة_مارك_روبنز	en la era de la información, la ignorancia es la elección de #markrobbinsnetwork.		
18	مبروك لصديقتنا في مدرسة فوغان الثانوية عشانها فازت لـ٩ مره بلقب جمعية يورك الرياضية . بالتوفيق...	felicitaciones a nuestra mejor amiga de la escuela secundaria de fugan por ganar el título de la club jurassic por novenas veces. buena suerte...		

Figure 5.4: **A sample of generated translations by our proposed NMT models using the transitive translation approach. The translations are from three models: French to/from Arabic-Levantine, Spanish to/from Arabic-Levantine, and Spanish to/from Arabic-Gulf.**

5.4 Multi-Lingual-Dialectal Online Social Behavior (OSB) Modeling

This section presents the methodology followed in modeling and evaluating our proposed framework for multi-lingual-dialectal online social behavior analysis.

5.4.1 Methodology

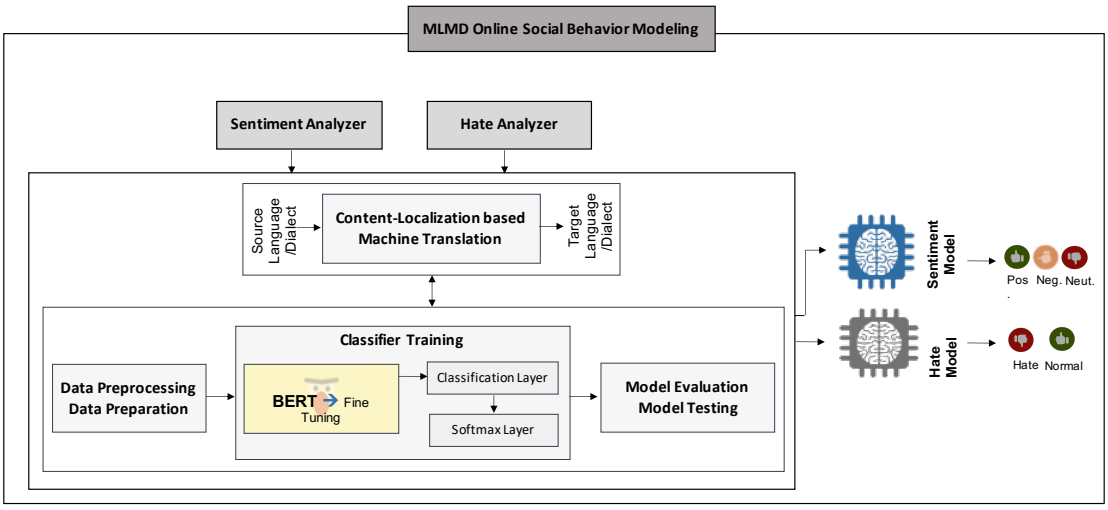


Figure 5.5: **Multi-lingual-dialectal online social behavior framework.**

Figure 5.5 illustrates the framework proposed to model multi-lingual-dialectal online social behaviors [21].

The framework depicted in Figure 5.5 is proposed as an attempt to minimize the dependency of languages and dialects in modeling online social behaviors. This framework is designed in a way that reduces the cost of building resources (i.e. data and models) for every language and dialect to perform OSB-related tasks by allowing exploiting existing resources (i.e. data and models) between low-resource and high-resource languages. To achieve this, we propose translating OSN data, namely messages, using content-localization approach (more details are discussed in Section 5.3. This way the context of contents is transferred to the language and even to the dialect of interest; hence, we can exploit existing resources to solve the problem in hand without building the needed resources (i.e. data and/or models) from scratch. This can be done in two scenarios: (1) localizing the contents of an existing data resource (i.e. collected, cleaned and filtered, and annotated dataset) to a low-resourced target language and/or dialect and using the translated content to train and build a model of the language and/or dialect of interest, (2) localizing the data of interest from a low-resource language and/or dialect to a high-resource language and/or dialect and using the already existing resource (i.e. model trained in a high-resourced

language and/or dialect) to solve the problem of interest. According to the literature, analyzing online social behavior in its native language produces better analysis, hence, we opt to use the first scenario to conduct the OSB analysis in this thesis. Further, deep learning approach is adopted for the training of online social behaviors and that has been covered and explained in details earlier in Chapter 4

5.4.2 Experimental Results & Analysis

In this section, we examine our multi-lingual-dialectal approach for solving the analysis of online social behavior tasks in different languages and dialects.

5.4.2.1 Experimental Design and Evaluation Protocol

We design our experiments to answer the following question: **can we exploit existing resources from a high-resourced language and/or dialect to solve an OSB problem in a low-resourced language and/or dialect using content-localization translation approach?**

We propose two experimental scenarios to answer the above-mentioned question:

- We translate an existing annotated dataset (i.e. source language/dialect) for a social behavior (i.e. sentiment in this experiment) using our proposed NMT models, into a target language/dialect. Note that we preserve the source annotations of the source dataset. Then, we train an OSB model using this translated dataset. Finally, we evaluate the trained models on external datasets under the condition that the contents of the external datasets should be in the native target language and/or dialect. English to Arabic-Levantine, English to Arabic-Gulf, and French to Arabic-Levantine are chosen as case studies for our proof of concept. In this experiment, we translate the English sentiment dataset SemEval 2013/20 [154, 182] to Arabic-Levantine and Arabic-Gulf dialects, using our proposed NMT models. The translated dataset will be used later to train Arabic sentiment classifier for Levantine and Gulf dialects. The same process is applied to the French sentiment dataset (FTSAD) [89]; it is translated to Arabic-Levantine dialect to be used later for training an Arabic-Levantine sentiment classifier. The purpose of this experiment is to examine the validity of our proposed approach across languages and dialects.

- We translate an existing annotated dataset X (i.e. from one source language/dialect) for a social behavior (i.e. toxic speech in this experiment) using our proposed NMT models into two target dialects, *aandb* of the same language Y (i.e. Arabic in this thesis). Note that we preserve the source annotations of the source dataset. Then, we train two OSB models using these translated datasets, one for each dialect (datasets d_{Ya}, d_{Yb}). After that, we choose an external dataset of dialect Y_a. Finally, we evaluate the trained models (i.e. models m_{Ya}, m_{Yb}) on the external dataset of dialect Y_a. The contents of this external dataset have been written and annotated in its native dialect

by native speakers Y_a. English to Arabic-Levantine, English to Arabic-Gulf, Spanish to Arabic-Gulf, and Spanish to Arabic-Levantine are chosen as case studies for our proof of concept. In this experiment, we translate the English hate speech dataset Hateval [33] to Arabic-Levantine and Arabic-Gulf dialects, using our proposed NMT models. The same process is applied to the Spanish hate dataset [145]; it is translated to Arabic-Levantine and Arabic-Gulf dialects to be used later for training an Arabic-Levantine and Arabic-Gulf hate classifiers. The purpose of this experiment is to investigate the effect of different dialects of the same language on the modeling and analysis of the online social behaviors on social media.

The translated versions of datasets (i.e. translated SemiEval 2013/2017, translated French FTSAD, translated Hateval, translated Spanish HateSpeech, to Arabic dialects using our proposed NMT models) have been used for training and validating sentiment and hate models, respectively. The datasets have been randomly split into 80% for training and 20% for validation. We evaluate our models on the validation set during training- every 100 steps- to track its learning progress. We have implemented early-stopping approach to regularize the learning of the model during training in order to prevent any potential overfitting to the training data. For evaluation measures, we have used accuracy, precision, recall, and F-Score, commonly used for classification evaluation, as evaluation metrics. Precision, recall and F-Score give a better view of model performance than accuracy alone does.

5.4.2.2 Datasets

We list below the datasets we have used for modeling and evaluating our proposed approach for multi-lingual-dialectal online social behavior analysis.

- **SemEval-2013/2017 Sentiment Dataset [154, 182]**: We have combined the sentiment English datasets of SemEval 2013 and 2017, and removed the neutral class. This has resulted in $\approx$ 15,000 tweets. We, then, have balanced the classes until we finally have a total of 11,024 tweets, out of which 6,500 are positive tweets and 4,523 are negative tweets.
- **French Twitter Sentiment Analysis Dataset (FTSAD) [89]**: This dataset contains 1.5 million French tweets translated from English to French. We have randomly sampled 20,000 tweets with a balanced distribution between positive and negative classes.
- **Saudi Banks Dataset [9]**: This dataset contains Arabic-Saudi tweets from four Saudi banks. The dataset has been manually annotated by two people. The dataset contains 8,669 negative tweets and 2,143 positive tweets.
- **Saudi Vision-2030 Dataset [15]**: This dataset contains tweets discussing several aspects of Saudi Vision 2030. The manually annotated tweets have yielded 2,436 positive tweets and 1,816 negative tweets.

- **ArSentD-Lev [30]**: This dataset consists of 4,000 tweets collected from the Arabic Levant region. The dataset has been manually annotated through the crowd-sourcing approach. There are 1,232 positive tweets and 1,884 negative tweets.

- **OCLAR datasets [165]**: OCLAR is an opinion corpus for Lebanese Arabic reviews. It contains customer reviews on restaurants, hotels, hospitals, etc., collected from Google Maps and Zomato website. The positive class contains all reviews rating from 1 to 3 (3,465 reviews), while the negative class contains the reviews with rating values from 1 to 2 (451 reviews).

- **HatEval Dataset [33]**: This English dataset has been constructed based on women or immigrants as targets of hate speech. The tweets have undergone two steps: (1) tweet annotations by non-expert annotators using crowd sourcing mechanism, (2) then two domain-expert annotators reviewed the annotated tweets. The inter-agreement in annotating the dataset scored 83%. The dataset contains a total of 13,000 tweets, out of which 5,470 tweets are labelled hate speech.

- **Spanish HateSpeech Dataset [145]**: This dataset is a subset of a bigger multi-lingual hate speech dataset that consists of 13 languages, one of which is Spanish. We extracted the Spanish samples. This resulted in a total number of 12,423 texts, out of which 4,239 are labeled as hate speech.

- **Let-Mi datasets [150]**: This dataset consists of Levantine tweets annotated for detecting misogynistic behavior on online social media. This dataset, which consists of 2,654 hate tweets and 2,586 non-hate tweets, has been annotated manually by Levantine people.

- **L-HSAB datasets [151]**: This dataset has been constructed and manually annotated for Levantine hate speech detection on social media. It contains a total of 5846 tweets, out of which 3,650 do not contain hate contents while 2,196 contain hate speech content.

- **COVID-19 datasets - Lebanon [235]**: This dataset has been collected from Twitter during COVID-19 pandemic in 2020. Geo-coordinates of Lebanon were used to collect and retrieve tweets in Arabic language during the COVID-19 pandemic. This dataset is used for evaluation purposes in our COVID-19 case study.

- **COVID-19 datasets - Saudi Arabia [5]**: This dataset has been collected from Twitter during COVID-19 pandemic in 2020. Geo-coordinates of Saudi Arabia were used to collect and retrieve tweets in Arabic language during eight consecutive days; 14-21 of March 2020. This dataset is used for evaluation purposes in our COVID-19 case study.

5.4.3 Results and Analysis

5.4.3.1 Performance of NMT-based OSB Models

The results in Table 5.7 presents the performance of the proposed OSB models that were trained using the translated dataset (i.e. translated from English and French to Arabic Levantine and Gulf dialects using our proposed NMT models). The results in Table 5.7 depict the models' performances using the validation set split of the same data that the models were trained on.

Table 5.7: **The performance of three sentiment models on the validation set in terms of accuracy, precision, and recall. The models represent English to Arabic-Levantine, English to Arabic-Gulf, French to Arabic-Levantine.**

		Validation Precision	Validation Recall	Validation F-Score	Accuracy
Model	En->Ar-Levantine	0.86	0.86	0.86	86
	En->Ar-Gulf	0.86	0.86	86	86
	Fr->Ar-Levantine	0.75	0.75	0.75	75

The English to Arabic-Levantine and English to Arabic-Gulf are shown to perform the same in terms of accuracy (86%), precision (0.86), and recall (0.86), followed by French to Arabic-Levantine model that scored 75% of accuracy, 0.75 points of precision and recall. There could be two reasons that the French to Arabic-Levantine models have shown a slightly less performance than the English-to Arabic models: (1) the NMT model of French to/from Arabic-Levantine dialect is trained on a subset of data compared to the English to/from Arabic dialects NMT models that are trained on the whole dataset (as explained previously in Section 5.3.4.4), (2) the pre-trained embeddings of large language models -that we have fine-tuned during our training- are different for each language; hence, this might have affected the learning performance of our French-Arabic OSB model. However, all the three models are shown to have effectively learnt the sentiment classes (i.e. positive and negative) using the translated dataset (i.e. by our proposed NMT models), and this is shown in the high performance metrics with at least 75% for accuracy, precision, and recall. Figures 5.6, 5.7, and 5.8 illustrate the high frequency words used for positive and negative classes for each model of the three. The words in the figures corresponding to the positive class (5.6(a), 5.7(a), and 5.8(a))-, reflect positive sentiment such as "متحمسة" meaning "excited", "طيب" meaning "good", "يحنن" meaning "so good", "مبسوطة" meaning "happy", "مرحبا" meaning "hello", "أحلى" meaning "better or more beautiful", "رائع" meaning "spectacular", "يضحك" meaning "funny", "أهلا وسهلا" meaning "welcome". Figures related to the negative class (5.6(b), 5.7(b), and 5.8(b)) also reflect negative sentiment such as "قتل" meaning "killing or murder", "غلط" meaning "wrong or mistake", "سيء" meaning "bad", "يضرب" meaning "beat", "غبي" meaning "stupid or dumb", "زعلان" meaning "sad or up-

set", "صعب" meaning "difficult", "مات" meaning "died", "أبكي" meaning "crying". From this, we can claim that our sentiment models are able to effectively learn sentiment from our translated data and to successfully distinguish between positive and negative classes which, in turn, proves the effectiveness of our multi-lingual-dialectal NMT models that have shown solid capability of transferring the context of OSN messages from language /dialect to another language/dialect.

(a) **Word-cloud for positive messages.**

(b) **Word-cloud for negative messages.**

Figure 5.6: **Word-cloud generated from messages classified as positive or negative sentiment by our English to Arabic-Levantine sentiment model.**

(a) **Word-cloud for positive messages.**

(b) **Word-cloud for negative messages.**

Figure 5.7: **Word-cloud generated from messages classified as positive or negative sentiment by our English to Arabic-Gulf sentiment model.**

(a) **Word-cloud for positive messages.**

(b) **Word-cloud for negative messages.**

Figure 5.8: **Word-cloud generated from messages classified as positive or negative sentiment by our French to Arabic-Levantine sentiment model.**

Table 5.8: **The performance of English to Arabic-Levantine, English to Arabic-Gulf, French to Arabic-Levantine, sentiment models, on external sentiment datasets, in terms of accuracy, precision, and recall.**

		F-Score (Positive)	F-Score (Negative)	Accuracy
Model	En -> Ar-Gulf evaluated on Saudi Banks Dataset [9]	0.76	0.93	89
	En->Ar-Gulf evaluated on Saudi Vision-2030 Dataset [15]	0.6	0.7	66
	En -> Ar-Levantine evaluated on ArSentD-Lev [30] + OCLAR datasets [165]	0.83	0.82	83
	Fr->Ar-Levantine evaluated on ArSentD-Lev [30] + OCLAR datasets [165]	0.81	0.7	77

While Table 5.7 presents the performance of the models using the validation set, Table 5.8 summarizes the performance of the models using external datasets. Each dataset that corresponds to a specific dialect is used to evaluate the model that has been trained on the same dialect (i.e. Levantine dataset is used to evaluate the Levantine model and the Gulf to evaluate the Gulf model). English to Arabic-Gulf OSB model is evaluated on two Saudi (i.e. Gulf dialect) sentiment datasets: Saudi Bank Reviews dataset [9], and Saudi Vision 2030 dataset [15]. The results show that the model is able to distinguish between both classes (i.e. positive and negative) in both datasets; it has performed at a positive f-score between 0.6-0.76, negative f-score between 0.7-0.93, and over all accuracy between 66%-89%. It can be observed that the English-Gulf model performs better on negative class than it does on positive class. After investigating the data, it has been found that there is mislabeling or ambiguity in some sentences as to whether they belong to the positive or negative class, as shown in the following examples:

- "للامانه انا مع الاهلي ليا عشر سنين وجيتهم هارب من سامبا واستخدمت جميع المنتجات من قروض شخصيه وعقاريه وبطايق فيزا واشوفهم افضل بنك ممكن تكون تجربتك الشخصيه سيءه ولكن لا تستعجل وتروح وتورط ممكن عشان الدمج فيه شويه لخبطه".
- "يستاهلون جميعا ساهموا في خدمه دينهم ووطنهم وليس من الانصاف اجحاف مجهودات رجال الامن لماذا لا تقدم لهم العروض كباقي منسوبي الصحه والتعليم والطيران".
- "يارب الوظيفة"

The mislabeling and ambiguity of labeling is a common issue found in existing datasets. This in fact affects the learning and evaluation process of online social behavior modeling. It is an ongoing challenge that labeling datasets, especially for online social behavior like sentiment, still puzzles the researchers in this area [125]. Overall, our English to Arabic

Gulf model has shown a reliable performance in detecting the online sentiment behavior. Below are three examples of a correct predicted sentiment examples:

- "الامير محمد بن سلمان يحقق رؤية المملكة ٢٠٣٠ بدعم الشباب".
- "هذا الرجل وضع للحق ميزان و بعدله استوى الأمير و الفقير أقنع بحربه على الفساد وب رؤية ٢٠٣٠ فيها المستقبل للوطن والعالم العربي والإسلامي و كافة الدول العضمى له منا الدعاء بأن يطيل الله في عمره على طاعته ليحقق ما نتطلع له من مستقبل باهر".
- "رؤية ٢٠٣٠ سوف تجعل السعوديه في مصاف الدول العضمى نثق برؤية سيدي".

A similar learning performance has been found in positive and negative classes from the English to Arabic-Levantine model on two Levantine sentiment datasets: ArSentD-Lev [30] and OCLAR [165]. The English to Arabic-Levantine model has scored a high learning accuracy of 88% with capability of separating positive and negative classes at positive and negative f-score of 0.83 and 0.82, respectively. Finally, the French to Arabic-Levantine has shown a competitive performance to the English to Arabic-Levantine model at classification accuracy of 77% with 0.81 f-score for positive class and 0.7 f-score for negative class.

5.4.3.1.1 Dialectal-Arabic Case Study: Sentiment Behavior Analysis during COVID-19 Pandemic in Lebanon and Saudi Arabia

Figure 5.9 illustrates the sentiment predictions of our NMT-based sentiment classifiers (i.e. Arabic-Levantine and Arabic-Gulf classifiers) on two COVID-19 datasets; one for Lebanon and the other for Saudi Arabia . Generally speaking, the sentiment tends to be more negative in Lebanon with 87% of negative sentiment than it is in Saudi Arabia with 65% of negative sentiment during COVID-19 pandemic in 2020. (as seen in Figure 5.9).

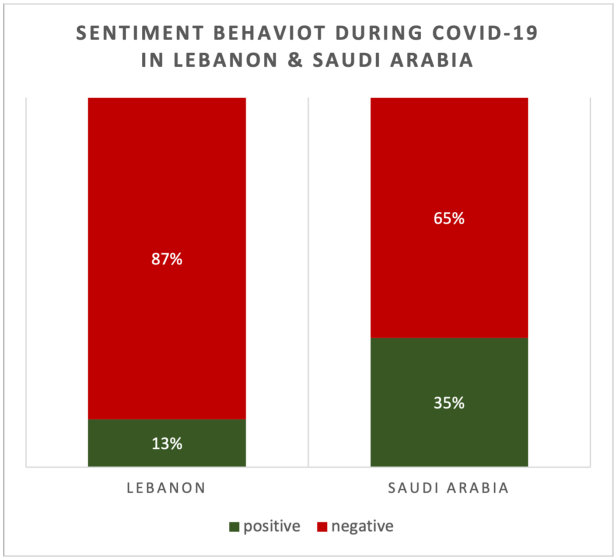


Figure 5.9: **Overall sentiment behavior in Lebanon and Saudi Arabia during COVID-19 Pandemic.**

Figure 5.10 provides a temporal view of sentiment behavior during the second week of March 2020, when COVID health measures were implemented in Saudi Arabia during COVID-19 pandemic. Overall, the sentiment is shown to decrease over time (as seen in Figure 5.10) especially after a positive spike on day 15 of March 2020. The temporal sentiment analysis for Lebanon is not provided in this case study due to the reason that date and time information is not available in the original Lebanon dataset [235].

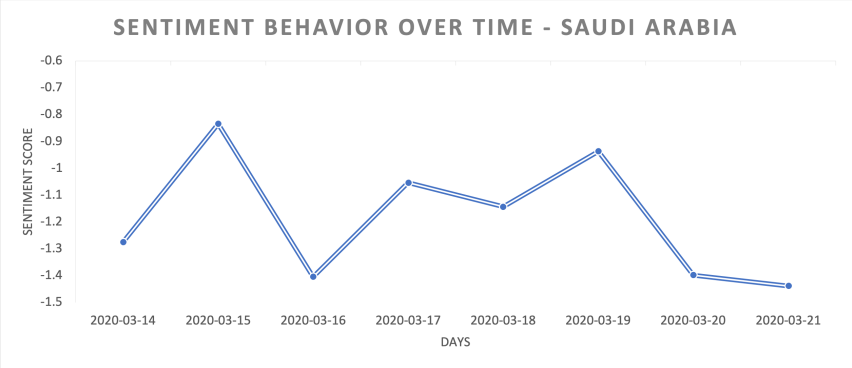
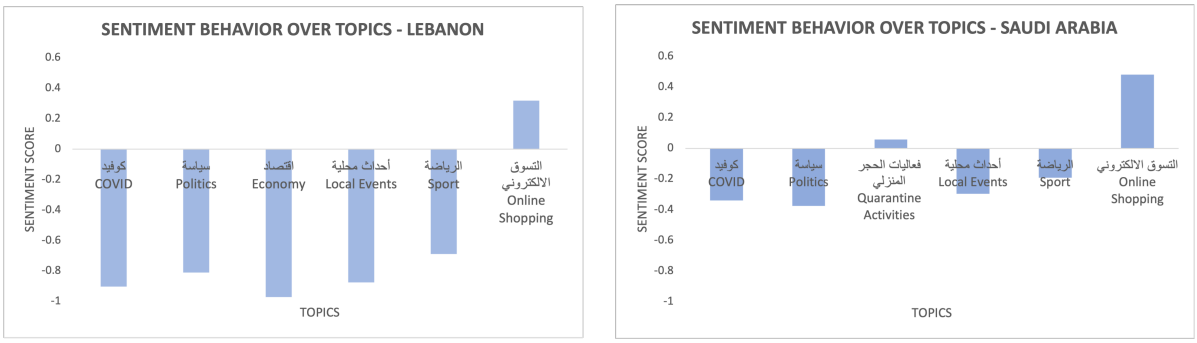


Figure 5.10: **Sentiment behavior over time in Saudi Arabia during COVID-19 Pandemic. The units are days according to the Saudi dataset [5].**

To facilitate the understanding of the sentiment behavior in Figures. 5.9 and 5.10, we provide a deeper analysis of the topics that people were discussing on OSNs platforms during the pandemic (Figure 5.11). Having the topics at hand, we are able to deduce the reasoning of the temporal abstract analysis. In other words, we can understand the reasons and causes of the inferred behavior. This has helped us to get a clearer picture of the story of events. The topics of both Lebanon and Saudi Arabia show an overall negative sentiment behavior with Lebanon having more negative vibes than Saudi Arabia does; while Saudi Arabia shows a more positive sentiment in two topics (i.e. quarantine/activities and online shopping), Lebanon shows a positive sentiment in only one topic (i.e. online shopping) only.



(a) (b)

Figure 5.11: **Comparisons of sentiment behavior between Lebanon and Saudi Arabia over inferred topics from COVID-19 data collected from Lebanon and Saudi Arabia during the pandemic.**

A wider exploratory analysis for the topic groups shown in Figure 5.11 is illustrated in Figures 5.12 and 5.13. The CVOID topic (Figure 5.11(a)) in Lebanon clearly shows a high negative sentiment, and the subtopics of the COVID topic shown in Figure 5.12, point out where the negative sentiment mainly come from. Subtopics like "virus", "China news", "John Hopkins", and "curfew" are shown to contribute to the high negativity in the COVID topic, while topics like "infected cases report" and "traditional treatment recipes (e.g. honey and lemon)" are shown to have positive sentiments. The politics topic in Lebanon has the highest negative sentiment among all other topics; the underlying subtopics of the politics topic explain the negative behavior by the residents of Lebanon toward international and regional news. It is interesting that the discussions on "Saudi Arabia" followed by "Kuwait" and "Jordan" subtopics, have shown a slight positive sentiment compared to the rest of other politics subtopics. Lebanon local events are discussed through subtopics shown in Figure 5.12(c), where subtopics "Hizb" and "Lebanon debates/news" show the highest negative sentiment followed by subtopics "citizens" and "private sector". The local subtopic "marriage", on the other hand, is the only subtopic among the other four that shows a majority of positive sentiment. Subtopics of economy and sport are shown to have a majority of negative vibes with sport showing a slight positive sentiment. The topic "online shopping" has an overall positive sentiment; however, its subtopic "order" is shown to have a negative sentiment compared to the "offers" subtopic that is shown to have high positive sentiments.

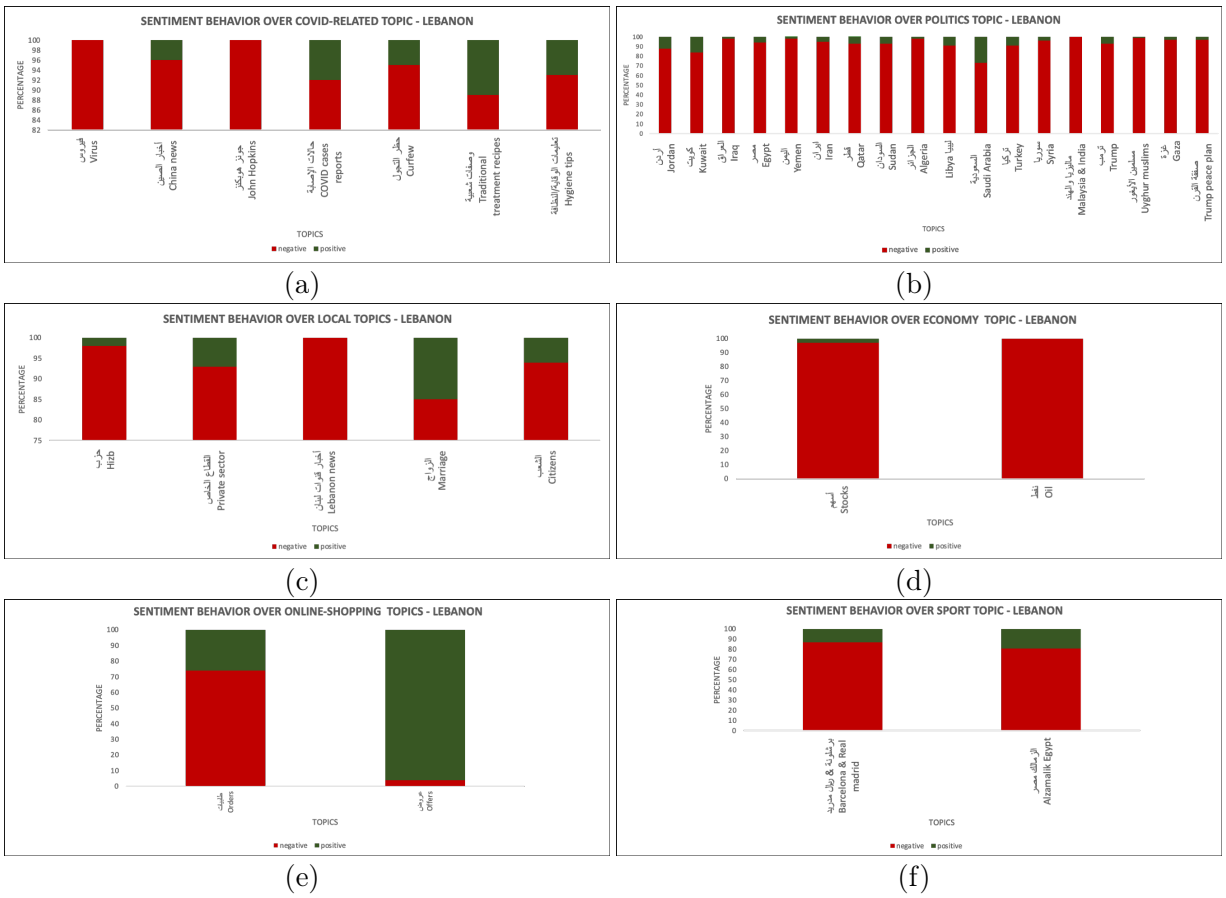


Figure 5.12: **Subtopics of the topics inferred from Lebanon COVID-19 dataset for sentiment behavior analysis in Lebanon.**

Taking a look at the subtopics of Saudi Arabia (Figure 5.13, we can notice that the people in Saudi Arabia are calmer and more relaxed than the people in Lebanon. It is clearly evident that the people in Saudi Arabia enjoy online shopping as it is used more often there than it is in Lebanon; many offers for "delivery services", "education services", "internet services offers", "discount codes" and "decor" can be observed in Saudi data during COVID-19 pandemic (i.e. in 2020) compared to only "discount codes" that were available in online shopping in Lebanon during COVID-19 pandemic in 2020. Moreover, People in Saudi seem to have had fun during the quarantine; the variety of activities' subtopics seen in Figure 5.13(d) show high positive vibes in most subtopics except for subtopics "mothers' situation" and "sleep and boredom"; the former expresses increased levels of mothers' frustration during the COVID-19 quarantine and the latter expresses the boredom during quarantining at home. The rest of the activities subtopic such as "tv series", "activities", "books and reading", "games", "music", and "reading holy Quran" illustrate high positive sentiment, especially in "activities" and "reading holy Quran" subtopics which show the highest positive sentiment. This observation reflects the religious spirit among Saudis, to whom Islam is a very important part of their life, and that all Saudi citizens and many of Saudi foreign residents are Islam believers. Many local Saudi topics are discussed through the subtopics in Figure 5.13(c). Subtopics such as "private sector", "students and dis-

tant education", "in-mosque praying hold", "hydro and bills", "Saudi banks" and "rents" show mostly negative sentiments. Subtopics that show positive sentiments are related to religion like "Ramadan", and support like "donation", "citizens' support", "thank-you to health workers", and surprisingly "quarantine". The positive sentiment associated with the "quarantine" subtopic reflects that the Saudi citizens seem to be willingly accepting and encouraged to respect the stay-at-home measure in order to prevent the spread of the corona virus. The sentiment of politics subtopics in Saudi Arabia (Figure 5.13(b)) seems to be more relaxed than its corresponding in Lebanon (Figure 5.12(b), where positive sentiments are barely seen. Sport topic has shown to have negative sentiment in both Lebanon and Saudi Arabia. The same negative sentiment has been found in sport topic in USA and Canada during the period of December, 2019 to April, 2020.

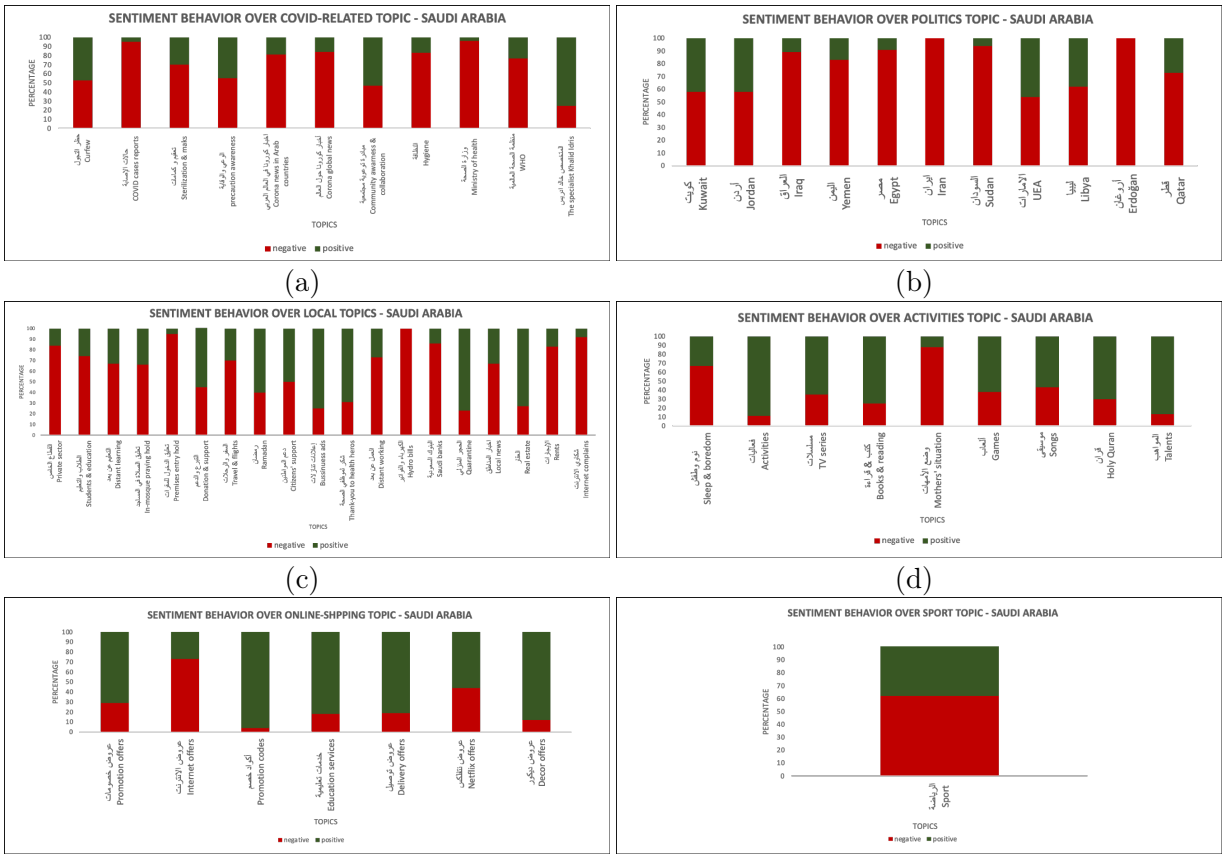


Figure 5.13: **Subtopics of the topics inferred from Saudi Arabia COVID-19 dataset for sentiment behavior analysis in Saudi Arabia.**

5.4.3.2 Cross-Dialect Performance of NMT-based OSB Models

Table 5.9 lists the performance results of our proposed hate OSB models that are trained on the translated datasets (i.e. translated from English and Spanish to Arabic Levantine and Gulf dialects, using our proposed NMT models). The results depicted in the table present the models' performances using the validation set split of the same data that the

Table 5.9: **The performance of four hate models on the validation set in terms of accuracy, precision, and recall. The models represent English to Arabic-Levantine, English to Arabic-Gulf, Spanish to Arabic-Levantine, and Spanish to Arabic-Gulf.**

		Validation Precision	Validation Recall	Validation F-Score	Accuracy
Model	En->Ar-Levantine	0.72	0.72	0.72	72
	En->Ar-Gulf	0.73	0.73	0.73	73
	Es->Ar-Levantine	69	0.68	0.68	69
	Es->Ar-Gulf	0.69	0.7	0.7	71

models have been trained on. As seen in Table 5.9, all models (i.e. The English-Arabic and Spanish-Arabic hate OSB models) have yielded solid performances in terms of accuracy between 69%-73%. The overall precision and recall for both hate and non-hate classes are quite high (i.e. at least ≈ 0.7 scores), which indicates that the models have efficiently learnt representative features that are able to detect the hate content correctly from the data.

Table 5.10: **The performance of English to Arabic-Levantine and English to Arabic-Gulf hate models on an external hate speech dataset in Levantine dialect, in terms of accuracy, precision, and recall.**

		En –> Ar-Levantine evaluated on Let-Mi dataset [150]						
		Hate			No-Hate			
		Precision	Recall	F-Score	Precision	Recall	F-Score	Accuracy
Model	En->Ar-Levantine	0.76	0.67	0.71	0.71	0.79	0.75	73
	En->Ar-Gulf	0.81	0.38	0.51	0.6	0.91	0.72	65

In Tables 5.10 and 5.11, we take a step further to assess our models on real datasets that have been constructed and annotated manually in a native language and dialect; the Levantine L-HSAB dataset, whose content is native Arabic Levant, has been collected from Levant geo-regions. For English models, it is clearly shown that the English to Arabic-Gulf model is not able to detect hate (i.e. or toxic) expressions in Arabic-Levantine dialect (i.e. messages in Let-Mi dataset [150]); While English-Gulf model has been able to recall only 38% of the hate contents, the English-Levantine model has been able to efficiently recognize hate contents expressed in its native dialect and recall 67% of the hate messages at as high precision as 76% while maintaining its ability to separate non-hate contents at a high performance as well (i.e. 71%, 79% for precision, recall, respectively). Similarly, the Spanish-Gulf model has shown a lower capability of detecting hate contents (i.e. toxic) expressed in a different dialect (i.e. Levantine) than its own dialect (i.e. Gulf). The Spanish-Gulf model has been able to recall only 54% of hate messages compared to the Spanish-Levantine model that has recalled 67% of hate contents at a precision score of 66%

Table 5.11: **The performance of Spanish to Arabic-Levantine and Spanish to Arabic-Gulf hate models on an external hate speech dataset in Levantine dialect, in terms of accuracy, precision, and recall.**

Model	Es –> Ar-Levantine evaluated on L-HSAB dataset [151]						
	Hate			No-Hate			
	Precision	Recall	F-Score	Precision	Recall	F-Score	Accuracy
Es->Ar-Levantine	0.66	0.68	0.67	0.8	0.79	0.8	75
Es->Ar-Gulf	0.7	0.54	0.61	0.76	0.86	0.8	74

while simultaneously keeping a high performance in distinguishing non-hate messages with as high as 80% of f-score of precision and recall.

Below we list example sentences expressed in Arabic-Levantine dialect, which the English-Gulf and Spanish-Gulf models have classified as non-hate contents, while our English-Levantine and Spanish-Levantine models have been able to classify as hate (i.e. toxic) contents.

- "@USER @USER اي نحن ما منقبلها صرماية بإجرنا ..مبروك. ع راسكم".
- "@USER انشالله بيقبر قلبك عن قريب ..يافهيمة عصرك".
- "@USER ضبي لسانك احسنلك".
- "@USER انشالله بيقبرك إنتي وعيلتك".
- "ما سمعنا جعاركم مبارح على القصف الإسرائيلي يلي اغتصب سماء سوريا ، نباح و عويل جيش ابو شحاطة وصل لقبرص".
- "ستموت بحسرة الخواجه ولن يعطيك شرف التحدث معه وابتزازه ايها الازعر السفيه ضبضب".
- "ليش بتنطرب وبتسكر لمن بتجيب سيرة الحذاء يا زلمي مع إن كل واحد بفش خلقو ع ذوقو حسب مفهومو ثقافتو وعقلو".
- "اكل الهوى للواطي دواء الله يشفيك".

As seen in the examples above, the same language has got different localized dialects; an expression in a certain dialect means something else in another and it is used in a different context. Ignoring such a feature can negatively impact the learning models so much that they end up generating misleading outputs. For instance, the Levantine idiom "صرماية بإجرنا -which means "shoes in our foot"- is a very local expression that Levantine people use in a negative situation (i.e. usually when in anger and it is used for swearing); however, the Gulf people do not use this idiom with the same structure; its equivalence though can be "شبشب في رجلي" or "نعال في رجلي". The same applies to the toxic expressions "بيقبرك، بيقبر قلبك، ابو شحاطة، ضبي لسانك، ازعر" - literally translated as

"bury you, bury your heart, father of thong-sandal, hold or fold your tongue, troubling person"- are used exclusively by Levantine people to express anger or dissatisfaction.

This finding highlights the importance of distinguishing dialects of the very same language and their localized contextual meanings. Overlooking those differences results in inaccurate understanding of the target dialect, which in turn leads to misleading and imprecise analysis of online social behaviors.

5.4.3.2.1 Dialectal-Arabic Case Study: Hate Behavior Analysis during COVID-19 Pandemic in Lebanon and Saudi Arabia

We have utilized our NMT-based Arabic-Levantine and Arabic-Gulf hate classifiers to analyze the hate speech behavior during COVID-19 pandemic in Lebanon and Saudi Arabia. Figure 5.14 shows the overall hate behavior detected in Lebanon is $> 2x$ greater than the hate behavior detected in Saudi Arabia.

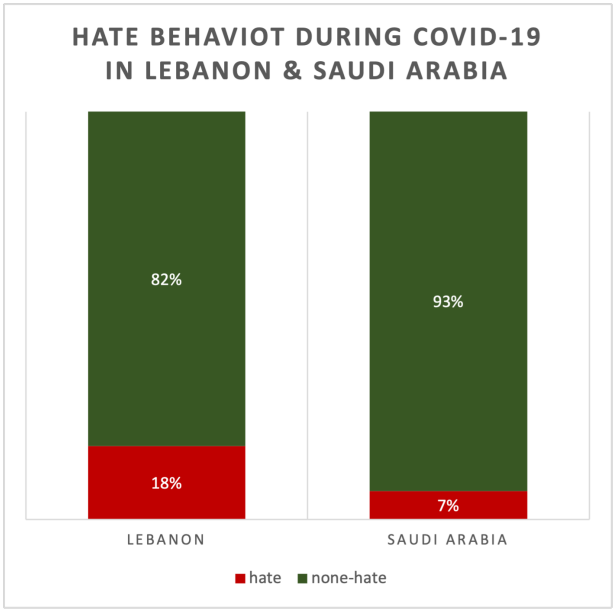


Figure 5.14: **Overall hate behavior in Lebanon and Saudi Arabia during COVID-19 pandemic.**

Figures. 5.15 and 5.16 demonstrate an exploratory analysis of hate behavior in both Lebanon and Saudi Arabia. We provide a detailed analysis from two views: temporal and topic-based analytic views. Figure 5.15 illustrates online social hate behavior as a timeline from Day 14 till Day 21 of March 2020. This period was when the announcements of restriction measures took place. From the figure, we can see that the hate behavior has slightly increased over time till it hits two spikes on days 20 and 21. On day 19 of March 2020, Saudi Arabia implemented a number of restriction measures such as in-mosque praying suspension. On day 20, March 2020, all domestic flights and train trips were suspended, and all shopping stores and supermarkets were to close from 8:00pm till 6:00am. Those series of measures explain the hate spikes in Saudi Arabia on days 20 and

21 of March 2020 during the COVID-19 pandemic. Note that the temporal hate analysis for Lebanon is not provided in this case study due to the reason that the date and time information is not available in the original Lebanon dataset [235].

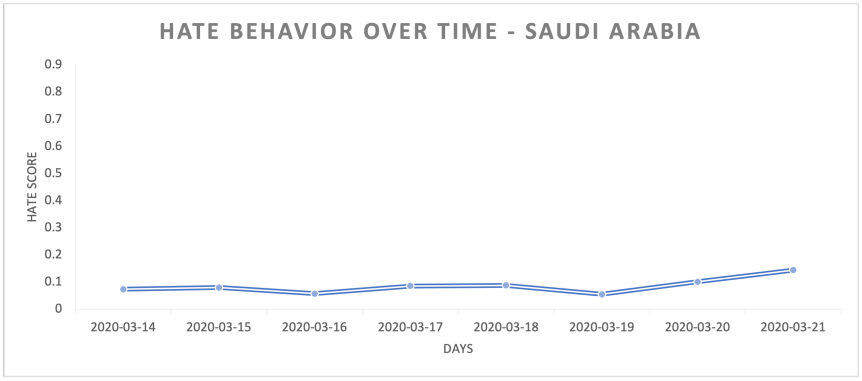
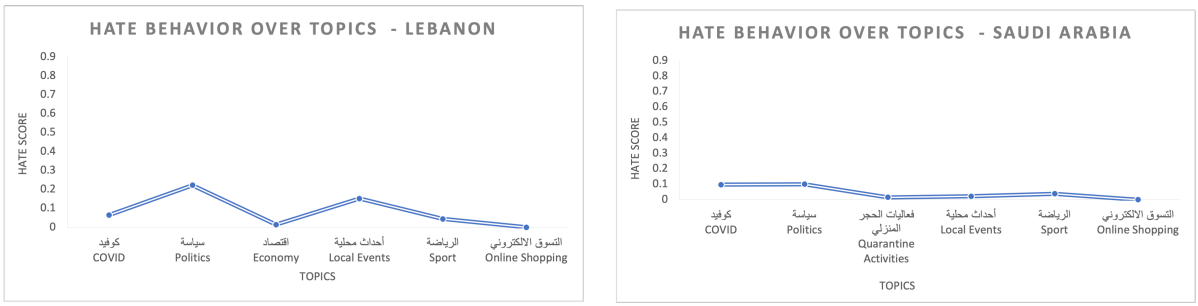


Figure 5.15: **Overall hate behavior in Saudi Arabia during COVID-19 Pandemic.**

As observed earlier, hate behavior in Lebanon is higher than that in Saudi Arabia during COVID-19 in 2020, especially in topics related to politics and local events, where the hate behavior scores the highest (Figure 5.16. Saudi Arabia show a slight hate behavior only in topics related to COVID-19 and politics; however, the detected hate behavior in Saudi Arabia is still $\approx 2x$ lower than that in Lebanon. COVID topic in Saudi Arabia (Figure 5.16(b)) is slightly higher in hate score compared to its corresponding in Lebanon (Figure 5.16(a)); this indicates that the people in Saudi Arabia seem to have been more upset about the virus spread and its consequences. "Sport" topic in both Lebanon and Saudi Arabia shows an insignificant level of hate behavior, while topics "economy", "quarantine activities", and "online shopping" score zero hate behavior in both Lebanon and Saudi Arabia.



(a) (b)

Figure 5.16: **Comparisons of hate behavior between Lebanon and Saudi Arabia over inferred topics from COVID-19 data collected from Lebanon and Saudi Arabia during the COVID-19 pandemic.**

Figure 5.17 shows the hate behavior detected in Lebanon subtopics associated with the topic categories depicted in Figure 5.16(a). The subtopic "curfew" in COVID topic (Figure 5.17(a) has a high hate behavior compared to the other subtopics; this subtopic is actually

responsible for making the hate behavior in COVID topic noticeable. Politics topic, which has the highest hate score during the COVID-19 pandemic in Lebanon, discusses eighteen subtopics (Figure 5.17(b)), out of which are "Iran", "Egypt", "Uyghur Muslims", and "Trump peace plan". Those subtopics show the highest hate behavior in the politics topic. During the COVID-19 pandemic in 2020, hate behavior has been detected in discussed local topics such as "Hizb which refers to Hizb Allah", "private sector" followed by "marriage and relationships" in Lebanon as seen in Figure 5.17(c). An insignificant hate behavior has been detected in "oil" subtopic (Figure 5.17(d)) associated with economy and "Alzamalik Egypt" subtopic associated with sport (Figure 5.17(f)). Zero hate behavior has been found in subtopics related to online shopping (Figure 5.17(e)). It is interesting to observe that Trump related subtopic of politics (Figure 5.17(b)) has zero hate score; this is not the case with "Trump" topic discussed in USA and Canada, where the hate score is high. The low hate behavior of "Trump" topic in Lebanon is due to the fact that most of the tweets talk about Trump report news, unlike the North America's tweets that seem to be opinionated against "Trump" by American and Canadian citizens and residents.

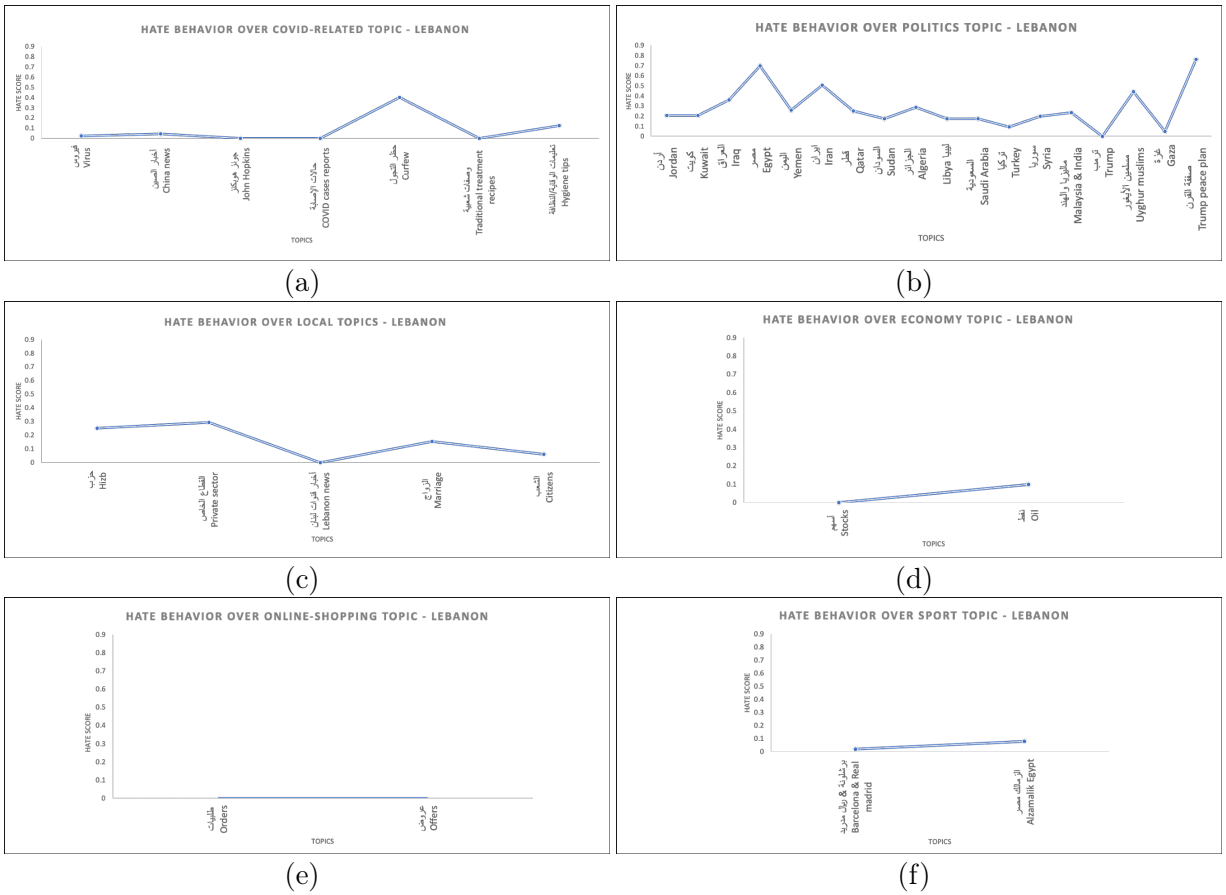


Figure 5.17: **Subtopics of the topics inferred from Lebanon COVID-19 dataset for hate behavior analysis in Lebanon.**

Generally speaking, online hate behavior is shown to be lower in Saudi Arabia and that is clear in the subtopics associated with the topic categories depicted in (Figure

5.18. Half of the topics and their subtopics report insignificant amount of hate presence and some report zero hate presence (Figures 5.18(d), 5.18(e), and 5.18(f)). The "curfew" subtopic of COVID topic has shown a spike in Saudi Arabia (Figure 5.18(a), which exhibits people's anger after implementing the quarantine restriction. Another COVID related subtopic "sterilization and masks" shows a hate behavior and this is due to the masks and sanitizers being insanely expensive or out of stock. Another presence of hate behavior found in politics subtopics especially in "Iran" and "Erdogan" (Figure 5.18(b)) scores the highest hate level among the other politics subtopics. Throughout the local topics discussed in Saudi COVID-19 data, subtopics "private sector", "in-mosque praying hold", and "internet complaints" show a slight amount of hate behavior (Figure 5.18(c)). The hate behavior detected in "private sector" reflects peoples' complaints about private sector being late in implementing restriction measures during the pandemic in Saudi Arabia. Religion is an essential part of life in Saudi Arabia; people go to mosques five times a day to perform five prayers every day. Therefore, implementing the measure of closing down mosques during the pandemic in Saudi Arabia was a major frustration to many, and that might explain the slight presence of hate behavior in the associated subtopic. The hate behavior detected in "internet complaints" subtopic is expected; during the pandemic all schools, universities, and work shifted to online, and people were instructed to stay home; however, some suffered from internet connection issues, which is most likely the cause of hate behavior.

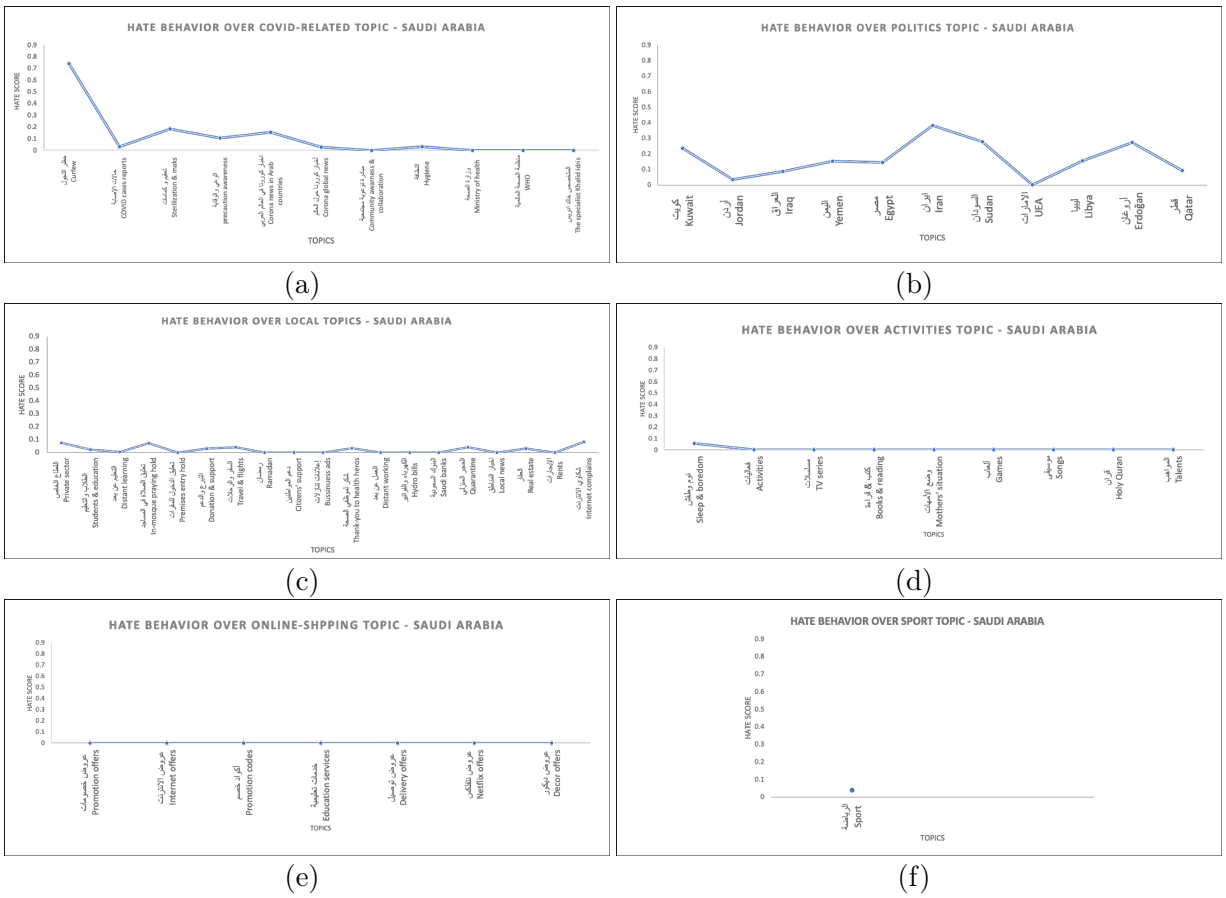


Figure 5.18: **Subtopics of the topics inferred from Saudi Arabia COVID-19 dataset for hate behavior analysis in Saudi Arabia.**

5.4.3.3 COVID-19 insights between North America and Middle East

A large-scale analysis on COVID-19 data in North America and Middle East has been conducted in this thesis as a proof of concept of our framework. Our exploratory analysis on COVID-19 data shows that the overall sentiment has been negative in USA, Canada, Lebanon, and Saudi Arabia. However, hate speech is shown to be the highest in Lebanon during COVID-19 pandemic; this is not due to the pandemic alone but it is also a protest against the wretched economic situation over there. Further, People in the Middle East are shown to be more interested in talking about international politics, with Lebanon showing a higher interest and more aggressive language usage when discussing politics than the case in Saudi Arabia. On the other hand, North America (i.e. USA and Canada) seems to have been concerned over the political issues related to the north continent like elections. "Trump" related topic seems to have opinionated tweets in North America, while it has been mostly news reports in the Middle East region. In addition, our analysis reveals a religious culture in Middle East through communicating prayers like "oh lord protect us and help us", encouraging worship practices on social media like reading the Quran and studying the Hadith, and expressing concerns over some restriction measures related to religion like the closure of mosques during the COVID-19 pandemic. This is barely the

case in North America where a small portion of messages have included religious content. NBA, hockey, and wrestling are shown to have been popular sports in North America while soccer is shown to have been the premium sport in Middle East. Fans of all sports in both continents have shown negative vibes mostly due to the sudden disturbance and cancellation of sport games during the pandemic. Public health seems to have been of concern for American and Canadian citizens, while local political issues have concerned Lebanese even during the pandemic. Saudis seem to have had local concerns related to religion (e.g. in-mosque praying hold) and services like internet and hydro during the pandemic. On the other hands, USA, Canada, Lebanon, and Saudi Arabia have shared the same concerns over distant learning and working from home as well as masks and sanitizers. Online shopping, promotions, and services have been shown to be used more in the Middle East (i.e. especially in Saudi Arabia) than it has in North America during the pandemic.

Chapter 6

Visual Online Social Behavior Modeling

6.1 Introduction

In this chapter, we explain our work on modeling the visual sub-component of the "Online Social Behavior Analysis" main component depicted in Figure 3.1.

People express their feelings, thoughts, and opinions in different visual ways, such as posting laughing faces, sunny beaches, or memes, on online social networks (OSNs), which contain diverse types of images. Images and visuals sometimes express feelings more clearly than words. On Twitter alone, tweets with images receive 89% more likes and 150% more retweets. In this work, we focus on exploiting different pieces of information within images. The aim is to analyze online social behavior (i.e., sentiment in this thesis) as a support tool for textual-based analysis or to understand online social behaviors on online social networks (OSNs) when textual-based tools are limited or even unavailable. One major challenge of developing systems for OSNs is their uncontrolled content; users have the freedom to populate data under open circumstances in an unstructured manner. OSN users have the control to start a trend or share a thought, which makes social media an open platform that is not restricted to any particular domain. This has raised a need to build flexible models and systems that are able to adapt or generalize to different domains -which is the main objective of this work. This introduces a technical challenge that arises from the lack of available datasets that can be used to train models that can be adapted or generalized to different domains. Researchers tend to construct datasets according to their specific domain of interest. We recognized this issue in our earlier study [19] and attempted to resolve it by proposing Domain Free-Multimedia-Sentiment Dataset (DFMSD). Our findings in [19] revealed that specific domain sentiment models do not generalize well on different domain-specific data. Unlike domain-specific models, general sentiment models have been shown to adapt well to domain-specific sentiment datasets. Note that our previous models (i.e. in previous chapters) were designed for textual tweets only. In this chapter, we investigate domain-independent sentiment extraction from visual content and conduct experimental evaluations using our image dataset (DFMSD). The images were collected under criteria-free conditions for use in models that work in the wild. The dataset contains different types of images that are categorized into three types: images containing faces, images

containing text, and images containing no faces/text. Our assumption is that each image type contributes valuable information that is embedded within the images and can be used to enhance the learning of sentiment behavior on online social media. In this chapter, we use sentiment as an online social behavior case study since we have constructed our own visual sentiment dataset. The sentiment includes three classes: positive, negative, and neutral.

Researchers in previous studies proposed to establish a direct mapping between sentiment orientation in images and visual features to solve visual sentiment analysis problem on social media [125]. However, images are freely shared on social media and their sentiment semantics are indirectly driven by cognitive semantics. This means that the relationship between the diverse OSN images and their sentiment orientation is extremely complex. Therefore, the usage of low-level features on social media images introduces a challenge of a semantic gap [125]. Researchers have utilized deep learning methods for visual sentiment analysis [48, 109, 121, 233] to fill the semantic gap between the low-level features and sentiment orientation. Deep learning-based methods make the sentiment predictions more interpretable as it transforms visual low-level features into an abstract feature space in which they benefit the analysis of emotional semantic for visual content on online social media. Although deep learning-based approach has achieved some progress in visual sentiment analysis on social media, current studies have not sufficiently given enough attention to the objects in visual content [125]. This means that the process of visual perception is ignored while establishing the mapping between image pixels and sentiment orientation. This thesis addresses this gap and proposes to consider two types of objects in images for visual perception: (1) faces to extract facial emotions and (2) texts to extract sentimental hints. Further, the unreliability of sentiment annotations in existing datasets affects the quality of visual sentiment models and hence increases the difficulty of network training. In this thesis, we propose a domain-free multimedia sentiment dataset (DFMSD) [1] that follows a high quality and strict protocol for data collection and annotation.

6.2 Datasets

This section presents popular datasets used for textual and visual sentiment analysis and facial emotion recognition on big data, in addition to our new sentiment multimedia dataset (DFSMD) [1].

- **Twitter for Sentiment Analysis (T4SA)** [215]: T4SA is a multimedia (i.e. texts and images) sentiment dataset that contains 1 million tweets with 1.5 million images. Texts were noisily annotated with three sentiment classes: positive, negative, and neutral. The images were annotated based on the sentiment associated with the texts. Due to the quality of the neutral class annotation observed during initial experiments, the model introduced confusion between the neutral class and both positive and negative classes. As a result, we removed the neutral class from TS4 dataset for our training purposes. T4SA dataset is used for the first stage finetuning of the basic visual sentiment modeling in this work.

- **AffectNet** [144]: AffectNet is the current largest facial emotion recognition in the wild dataset that was manually annotated. It consists of 1M facial imagesthat are greyscale with 48*48 pixels; 44K of which were manually annotated with eight facial emotions: neutral, happy, sad, surprise, fear, disgust, anger, and contempt. Since we study the sentiment in the wild on social media, there exit animated and cartoon images, especially faces in memes, for example. Therefore, we convert AffectNet images into animated version and combine those animated images with the original ones. The modified version of AffectNet dataset is used for the first stage finetuning of our facial emotion recognition modeling in this work. Note that we convert the images of AffectNet to greyscale since we use FER-2013 data for building our final FER model for this work.

- **FER-2013** [84]: FER-2013 is a well-known facial emotion recognition dataset that has been used extensively in modeling FER models and applications. It consists of 35K images that are greyscale with 48*48 pixels. The facial images were manually annotated with seven emotions: angry, disgust, fear, happy, sad, surprise, and neutral. FER-2013 dataset is used for building our FER model used in this work.

- **DFMSD** [1]: Domain free multimedia sentiment dataset (DFMSD) is our newly introduced dataset for visual and textual sentiment analysis. It was designed and constructed to work in the wild and to be able to deal with uncontrolled conditions in online social media. DFSMD was collected using Twitter Stream API. The protocol followed to collect and annotate DFSMD makes it distinguished from other datasets as data collection process was not restricted to any keywords, domains, locations, or any predefined retrieval criteria. The annotation questions and annotators of the dataset were selected carefully to minimize any possible biases during the annotation. Moreover, the annotators of the dataset were selected on the basis of providing sentiment agreement with three expert psychologists. The DFSMD consists of 14,488 tweets which contain 10244 images. 46% (i.e.6683 tweets) of the tweets are positive, 33% (i.e. 4822) are negative, and 21% (i.e. 2983) are neutral. The images distribution follows 47% belonging to the positive class, 10% belonging to the negative class, and 43% belonging to the neutral class. Note that texts and images were annotated separately in a way that the annotation of texts does not affect the annotation of the images. We decided to extend our sentiment image dataset by following the same collection and annotation approach used earlier as an attempt to improve the deep learning performance and to minimize the problem of a sever class imbalance. The first version was published in an earlier study [1]. Figure 6.1 illustrates a sample of our images annotated in positive, negative, and neutral sentiment classes.

(a) **Positive images.** (b) **Positive images.** (c) **Negative images.**

Figure 6.1: **A sample images, from our DFSMD dataset, that were manually annotated into positive, negative, and neutral classes.**

6.3 Method

This section presents in detail the method we followed to model our multi-model visual social behavior analysis; it includes data preprocessing, datasets, and all the approaches we adapted for our final model.

6.3.1 Preprocessing:

The step of preprocessing is very important for the learning process; we clean, denoise, prepare the data before we feed to the VIT training. The image preprocessing consists of the following steps:

- Face detection: we use the face detection algorithm proposed in [46] that uses facial keypoints to detect faces. The face detector finds four coordinates of region of interest (ROI) of faces then the detected face(s) is cropped, and all irrelevant background is discarded. Also faces that are far and not clear are discarded with respect to the ratio of the face and the image size. If the ratio is less than a predefined threshold we discard the face. with respect to an image and if their value is less than a predefined threshold. This step is applied to the facial emotion recognition modeling part.
- Color to greyscale conversion. This step is applied in the facial emotion recognition modeling part.
- Data augmentation: it increases the size of dataset and deal with imbalance between classes since deep learning works better with more data. We use 2 types of augmentation one for training and another for testing:
  - Training-time: Images are randomly resized in the range of (224,350), and then center-cropped by cropping size of (224, 224). Then the images are randomly flipped.

– Testing-time: Images are resized to (256,256), then cropped using 10-crop technique. 10-crop technique resizes an image to (256,256) and apply 5 crops (upper-left, upper-right, lower-left, lower-right, center) with a cropping size of (224,224) and then apply L-R flipping which will result in 10 cropped-flipped images. Finally, we use the average prediction of these 10 images [94]. This step is applied in basic visual sentiment modeling and facial emotion recognition modeling parts.

- Text detection and extraction: we use Optical Character Recognition (OCR) to detect and recognize texts from images. Extracted words are not in sorted order after OCR extraction, hence, we sort the extracted words in order of their occurrence using contours detection; to separate the different lines. Then, we simply process the contours left to right to sort the words within lines. Finally we apply quantization on heights with a pre-defined threshold to group words in the same line together. This step is applied in the textual-images part.

6.3.2 Visual Transformers Model

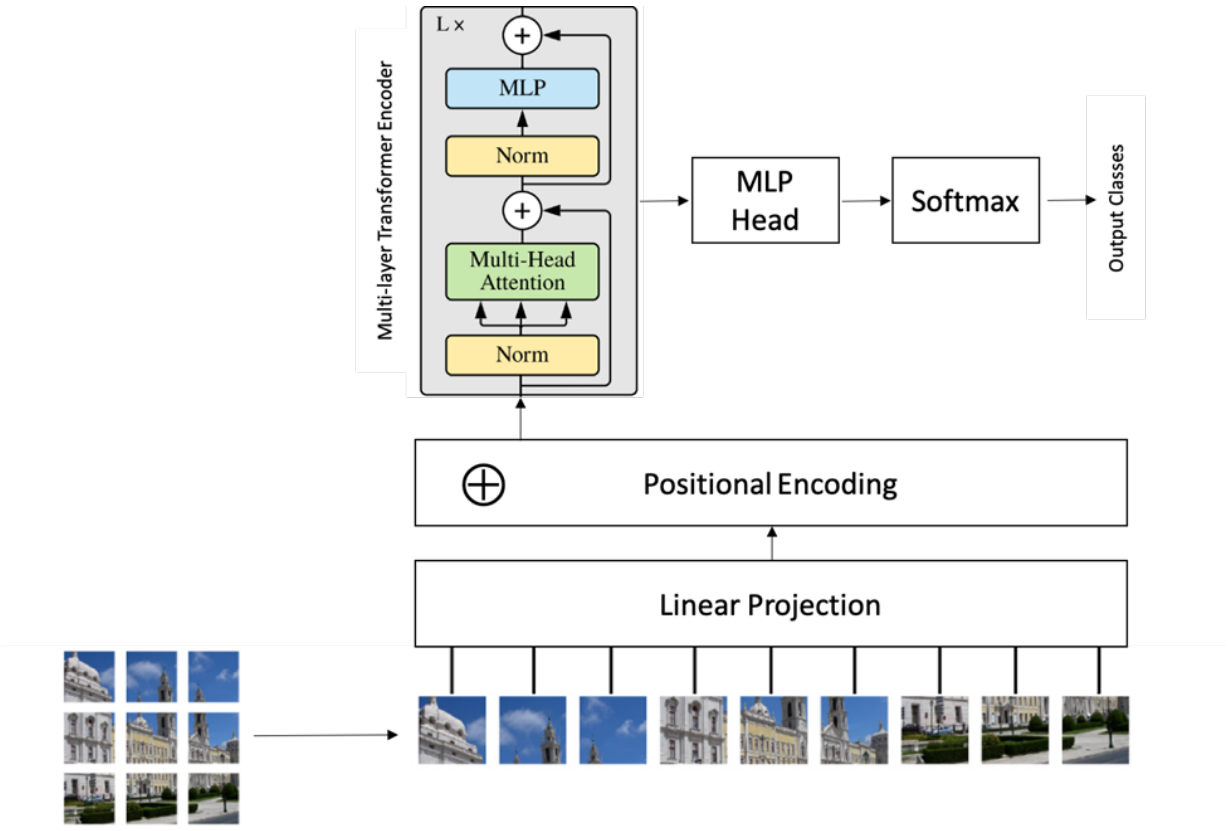


Figure 6.2: **The architecture of visual Transformers [62] used in training our models.**

Visual transformers (Vit) [62] – which was introduced recently in 2020- is used as the

deep learning architecture in our work. ViT has been a competitive alternative to CNNs for image recognition tasks. It outperforms the current-state-of-the-art CNNs by four times in terms of computational efficiency and accuracy [62] especially on big date regimes. In big data regimes, the inductive biases put in CNNs are not needed; instead ViT can learn those biases by itself. Shallower layers of ViT are able to localize attention (i.e. attend to local pixels) and globalize attention (i.e. attend to global pixels) compared to CNNs which only have local small receptive fields in shallower layers. The advantage of shallower layers being able to attend to local and global pixels in images is that it allows the ViT model to learn how and when to attend and the bias is not needed anymore. Because of its efficiency in handling big data regimes [62], it can be widely used in systems and applications related to online social networks -where data sizes are huge- with remarkable performance.

A high-level overview of the ViT architecture is given in Figure 6.2. An input image is split into patches and then flip the patches in order and flatten out them. A linear projection is applied to the flattened patches and then we add a positional encoding before we feed them to the multi-layer transformer encoder. The structure of a single encoder consists of a multi-head attention module and multi-layer perceptron module. The output of the encoder is then fed not a multi-layer perceptron head which is used a classification module that yields class predications.

6.3.3 Threshold-Moving

Data imbalance is a common phenomenon in sentiment datasets collected through OSNs [1, 19, 26]. This is also observed in our DFSMD dataset. Although the literature [226] suggests that a balanced dataset would improve model learning, it is too expensive to balance the data while simultaneously preserving the natural distribution to avoid biases. To overcome the class imbalance issue in this work, we propose fusing the threshold-moving approach with the sentiment learning process as an attempt to enhance the learning performance of our models. It has also been proven that the transformer architecture [17] is effective in dealing with class imbalance.

The default decision threshold (i.e. 0.5) for classification problems with class imbalance might negatively impact the learning performance and hence yield poor results. The default decision threshold might not represent the optimal interpretation of a model's predicted probabilities. As such, a simple approach to improve the classification performance on an imbalanced data is to tune this hyperparameter (i.e. threshold) that is used to map the predicted probabilities to class labels. The process of tuning this hyperparameter is called threshold moving. In this work, we calculate the optimal threshold using grid search approach. We search threshold values for a model and consider the best value that yields the best performance in terms of our evaluation metric. We apply threshold moving during the training process in such a way that in each validation iteration (i.e., epoch) we examine a range of threshold values on the predicted class probabilities in order to find the best threshold. The threshold that achieves the best performance (i.e. in terms of evaluation metric) is then adopted for the model (i.e. at the current iteration or epoch). Upon the completion of the training, the model with the best performance is chosen to make predictions on new data.

6.3.4 Two-Stage Strategy

Transformers have been shown to perform best when trained on large-scale datasets [62, 112]. This means that transformers are able to generalize well on classification tasks when trained on large-scale datasets compared to when trained on small datasets. With the limitation of existing datasets, which are small in size, researchers have exploited the transfer learning approach to benefit from transformer models pretrained on large-scale datasets. Given that it is expensive to curate large-scale datasets of high quality in terms of time, cost, and human labor, transfer learning offers a reliable solution for learning classification tasks using small datasets. To overcome the differences between source and target tasks, the strategy of two-stage learning (i.e., finetuning) has been recommended in the literature [140, 220] to compensate for the limitation of small datasets.

Transfer learning offers a rich set of benefits including improving the efficiency of model training and saving of time and resources since building a high-performance model from a scratch requires a large amount of data, time, resources, and efforts. Therefore, we use a two-stage learning approach [220] to train ViT models as an attempt to solve the limitation of small labeled datasets. We implement the first stage fine tuning on huge-size datasets to maximize the benefits of transfer learning. We use a ViT architecture with pretrained weights from ImageNet-21K dataset [62]. The whole ViT architecture was retrained for the first stage fine tuning. For the basic sentiment model, we use TS4 dataset to finetune the ViT pretrained model. The model learns two sentiment classes: positive and negative. To overcome the class imbalance between the two classes, we fuse the threshold moving approach with the sentiment learning process in this part. For the facial emotion expression model, we finetune the pretrained ViT model using a modified version of AffectNet dataset. The modified version of the AffectNet dataset includes its original images in addition to the same images converted into animated version. The model learns eight classes: neutral, happy, sad, surprise, fear, disgust, anger, and contempt.

For the second stage fine tuning, we initialize the ViT architecture with the weights obtained from our first stage fine tuning. The last fully connected layer is replaced by a new MLP head for both models: basic sentiment and facial emotion models. For the basic sentiment classifier, it outputs two classes: positive and negative. our DFSM dataset is used for the second stage finetuning with adopting the threshold moving approach during the process of sentiment learning on the positive and negative classes. The finetuning in the second stage is done through training the whole ViT architecture. We take the output of the last layer (i.e. high level representations of the mode) and feed it as features to our visual multimodal sentiment classifier. The facial emotion classifier outputs seven classes: angry, disgust, fear, happy, sad, surprise, and neutral. FER-2013 dataset is used to finetune the AffectNet pretrained model obtained from our first stage finetuning. The finetuning in the second stage is done through training the whole ViT architecture. We take the output of the last layer (i.e. high level representations of the mode) and feed it as features to our visual multimodal sentiment classifier.

6.3.5 Visual Deep Multi-Modality Fusion

The learning of multimodal sentiment recognition requires two components: a feature extraction method and a fusion strategy. In this work, we adapt the multimodality approach in [115, 169] and propose training three separate transformer-based deep models to extract features from three types of images (i.e., tasks): images containing faces, images containing text, images containing no faces/text. We adopt the intermediate fusion approach to fuse the extracted features and feed them to an MLP architecture to build our final multimodal online social behavior model (i.e. a sentiment model as a case study in this thesis). To the best of our knowledge, we are the first to use a transformer-based fusion approach with pretrained transformer-based models to extract features for multimodal sentiment analysis on OSNs. Additionally, we believe that we are the first to fuse the threshold-moving approach with the learning process using the transformer architecture.

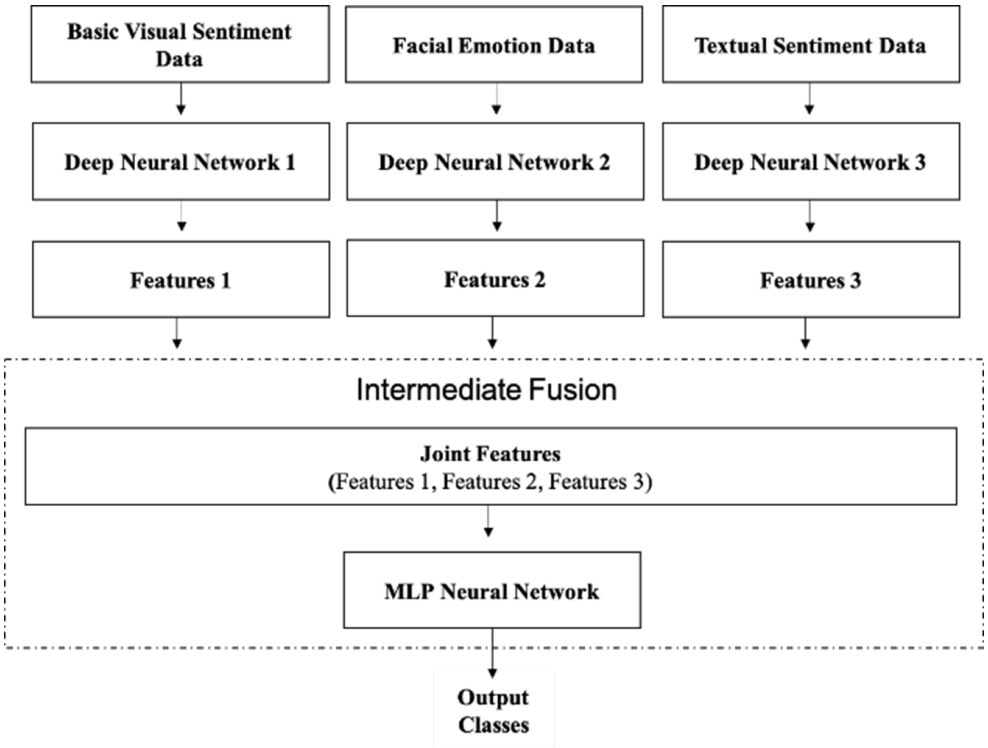


Figure 6.3: **The proposed architecture for ViT-based multi-modality fusion for visual online social behavior analysis.**

We use the architecture of intermediate fusion to fuse our deep learning based models with a goal of building a multi-modality online social behavior model. Figure 6.3 illustrates our proposed architecture for developing multi-modality online social behavior classifier (i.e. a sentiment case study in this work). The intermediate fusion allows for data fusion at different stages of model deep learning as it offers flexibility to fuse features at different depths. Deep learning based multi-modal data fusion has shown great improvement in learning performance [65, 169]. The input for the intermediate fusion is the higher-level representations (i.e. features) obtained through multiple layers of deep learning. Hence,

the intermediate fusion in the context of multi-modal deep learning is the simultaneous fusion of different model representations into a hidden layer so that the model learns a representation from each of the individual model. This layer where the fusion is done at is called a fusion layer. This work proposes a ViT-based fusion for a multi-modality visual online social behavior analysis. Three models -single-modality sentiment, facial emotion, textual sentiment) are trained using the backbone of Transformers (i.e. ViT for visual contents and BERT for textual contents). Then those models are used to extract deep features before all are fused to form one joint feature that will be fed into an MLP classification head.

6.4 Experimental Results and Analysis

This sections presents the experimental results and analysis for the visual models. The details of modeling the textual model is covered in Chapter 4.

the first-stage fine-tuning of ViT architecture was implemented with pretrained weights obtained from ImageNet-21K dataset for both single-modality sentiment and FER models.

6.4.1 Performance of Single-Modality Visual Sentiment Model

6.4.1.1 First Stage Finetuning

We implement the first-stage fine-tuning of ViT architecture using T4S4 dataset. Due to a poor quality annotation of the neutral class (i.e. after preliminary experimentation), we decided to train the model on the positive and negative classes. Table 6.1 shows the performance of our ViT-based single-modality sentiment model in the first-stage finetuning. Table 6.1 shows the results of the performance when fusing threshold-moving with the training. The threshold-moving has shown to absolutely enhance the learning performance by six points in term of accuracy, eight points in term of positive F-Score, and five points in term of negative class. This model will be used for image generic deep feature extraction in order to learn our multi-modality visual sentiment classifier.

Note that we present the learning performance (ie. without threshold moving applied) for three classes as well. The behavior of during the training has shown that the neutral instances confuses the model between both positive and negative classes.

6.4.1.2 Second Stage Finetuning

In Table 6.2, we demonstrate the performance of our second-stage ViT model using the images from our DFMSD dataset. Based on the first-stage pretrained model, The performance accuracy of of learning two classes (i.e. positive and negative) (i.e. with threshold-moving technique fused during the training) is 81% with F-Score scores of 0.86, 0.7 for positive and negative classes, respectively. Further, the results of second-stage finetuning has shown the

Table 6.1: **The performance of first-stage single-modality ViT sentiment model on T4SA dataset.**

	Precision			**Recall**			**F-Score**			**Accuracy**
	Positive	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**	
Binary classes	0.64	0.61	-	0.58	0.67	-	0.61	0.64	-	63
Binary classes with threshold moving	0.69	0.69	-	0.69	0.69	-	0.69	0.69	-	69
3 classes	0.51	0.48	0.47	0.37	0.54	0.55	0.43	0.51	0.51	49

effectiveness of the two-strategy finetuning approach; the learning performance has greatly improved by 12 points in term of accuracy, 29 points in term of positive F-Score, 8 scores for negative F-Score. We observe that the performance for the neutral class has decreased in term of precision whcih explains an introduced confusion with the other classes. Overall, the model performs well in distinguishing between positive and negative classes since it has high F-Score scores > 0.70 and recall scores $\geq \approx 0.60$ for both classes.

Table 6.2: **The performance of second-stage single-modality ViT sentiment model on images from our DFMSD dataset.**

	Precision			**Recall**			**F-Score**			**Accuracy**
	Positive	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**	
Binary classes with threshold moving	0.88	0.67	-	0.84	0.73	-	0.86	0.7	-	81
3 classes	0.77	0.66	0.36	0.67	0.54	0.55	0.72	0.59	0.44	61

6.4.2 Performance of Facial Emotion Model

To train all FER models, all faces have be to be detected and cropped as explained in the preprocessing section.

6.4.2.1 First Stage Finetuning

We implement the first-stage fine-tuning of ViT architecture using AffectNet dataset. The model achieved ana ccuracy of 59% with F-Score of 0.59 for 8 classes.

6.4.2.2 Second Stage Finetuning

Based on the weights obtained from the first-stage ViT finetuning using AffectNet, we implement the second-stage finetuning with the pretrained weights obtained from the first

stage, using FER2013 dataset. Table 6.3 illustrates that effectiveness of the two-stage strategy in enhancing the learning performance -in term of precision, recall, and accuracy- between multiple classes. This can be obviously observed in the recall of the model that has been greatly improved by 7 points when applying two-stage strategy compared to only using one stage for finetuning. This model will be used for facial emotion deep feature extraction in order to learn our multi-modality visual sentiment classifier.

Table 6.3: **The performance of second-stage ViT FER model on FER-2013 dataset.**

	Precision	**Recall**	**F-Score**	**Accuracy**
FER2013- one-stage	0.68	0.62	0.64	69
Fer2013- two-stage	**0.71**	**0.7**	**0.7**	**70**

6.4.3 Performance of Multi-Modality Visual Sentiment Model

Table 6.4 shows the effect of using the extra information of facial emotion and texts residing in images, in addition to the information of images themselves. Fusing ViT features of our FER and single-modality models has shown a slight improvement in the performance compared to the performance of the single-modality ViT sentiment model, in term of accuracy and F-Score for both positive and negative classes. While fusing single-modality sentiment and FER features noticeably improves the negative precision (i.e. by 6 points), it drops the recall of the positive class (i.e. by 4 points). However, fusing textual and facial emotion features along with the single-modality sentiment stabilizes the learning and further improves the overall performance for all the classes in term of F-Score. In more details, negative precision improves by 4 points without affecting the positive recall and similarly positive recall improves by 3 points without affecting the negative recall.

We further examine the effect of fusing facial emotion and text features with the single-modality sentiment on three classes: positive, negative, and neutral. It can be seen from Table 6.4 that the overall F-Score of the model improves -especially for the negative and neutral classes- when fusing the three types of features compared to only two.

Table 6.4: **Performance of fusing three types of ViT-based deep features extracted from three pretrained models: single-modality sentiment, facial emotion, and textual sentiment. The performance is evaluated in term of accuracy, precision, recall, and F-Score.**

	Precision			**Recall**			**F-Score**			**Accuracy**
	Positive	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**	
Single-modality Sentiment model (ViT)	0.88	0.67	-	0.84	0.73	-	0.86	0.7	-	81
MLP (Sentiment + FER ViT deep features) - 2 classes	0.86	0.73	-	0.88	0.69	-	0.87	0.71	-	82
MLP (Sentiment + FER + text ViT+BER Tdeep features) - 2 classes	0.88	0.71	-	0.87	73	-	0.87	0.72	-	82
MLP (Sentiment + FER ViT deep features) - 3 classes	0.73	0.66	0.43	0.77	0.54	0.44	0.75	0.59	0.44	64
MLP (Sentiment + FER + text ViT+BERT deep features) - 3 classes	0.75	0.64	0.42	0.73	0.57	0.5	0.74	0.6	0.46	64

Chapter 7

Topic Modeling and Dynamic Topic Interpretation

7.1 Introduction

In this chapter, we present our proposed "Topic modeling and Interpretation" component (Figure 3.1) that is responsible for developing real-time topic modeling and dynamic interpretation models customized for social media contents.

Give the continuous and heavy information flow during on social media, there is an urgent need to discover global and local trends, concerns, and issues in real time, yet it is nearly impossible to achieve this manually. The unsupervised learning nature of topic modeling methods makes it possible to achieve this goal fast and without prior human knowledge involved. However, these methods are sensitive to data noises, which is a normal phenomenon in OSNs data. It is well-known that social media data suffers from noises such as misspellings and the intense use of abbreviations due to the limited writing-space capacities. This challenge requires a careful handling of the OSNs data in order for the topic modeling methods to perform well. Given the issues of data noises and short lengths, proper techniques are needed to extract every informative piece of information from messages while synchronously removing the unnecessary noises residing within these messages. Although topic modeling methods are proven effective in capturing hidden insights from the social data, it is not capable of providing a coherent interpretation of the inferred insights [138]. However, studies [14, 232] still depend on using the top n words resulted from these topic models to interpret the topics (i.e. discovered insights or clusters) on social media. Another cheap alternative that has been considered is to interpret the topics manually. Manual topic interpretations requires human efforts and can be easily biased towards subjective opinions [136]. Both approaches are not applicable for pandemic-friendly systems which require accurate and instant interpretations in order to make proper decisions. According to the literature, people prefer phrases over single words to understand topics [138]. They claim that combining single words creates difficulties to comprehend the main meaning of topics while sentences are too specific and might miss other aspects of topics. Given the diversity of conversations on OSNs, finding the optimal

length of phrases that best describe a topic is challenging. Current methods rely on a fixed sliding window for phrases which might limit the comprehension of topics. Some topics might use longer or shorter phrase-expressions than the other and this cannot be controlled on open platforms like OSNs that encourage unstructured data format.

This work exploits the Natural Language Processing (NLP) techniques as pre-requirements to topic modeling in order to maximize the learning performance of topic models on OSNs. In addition, unsupervised learning approach is used to find phrases of dynamic sizes that provide coherent interpretations of the inferred topics automatically. We use the data explorations and interpretations as complementary tools to facilitate the understanding of online social behavior. This work is an attempt to assist in catering to public safety and psycho-social needs towards providing measures for developing healthy coping strategies to reduce the psycho-social instabilities for events like pandemics. It could also create opportunities for tracing individuals or groups responsible for violent incitement as it has been proven that it is possible to infer this type of information through OSNs [22, 184].

7.2 Background

This section presents a background overview of the methods used to design the proposed framework. The background is composed of three parts: topic modeling, phrase extraction, and deep sequence classification.

7.2.1 Topic Modeling

Topic modeling is an unsupervised learning technique that detects patterns of words and expressions within datasets, and automatically determines clusters of similar words and phrases that best characterize a set of texts. Recently, topic models have been increasingly used to explore and infer insights from social media data [107, 158, 162]. Both traditional learning and deep learning approaches have been used to learn hidden topics from social media data [3, 14, 107, 235]. Latent Drichelet Allocation (LDA) [37], Non-negative Matrix Factorization (NMF) [118], BERTopic [86] are examples of the most prominent algorithms for topic modeling in social media analysis.

7.2.1.1 Traditional Topic Modeling Approach

In this section, we provide an overview of two traditional topic molding algorithms; Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF).

7.2.1.1.1 Latent Dirichlet Allocation (LDA)

LDA [37] is a probabilistic generative algorithm that utilizes Bayesian framework and Dirichlet distribution. It treats a collection of data as a mixture of latent themes or topics,

where each topic is considered a multinomial distribution over a fixed vocabulary. LDA considers two matrices to determine the hidden patterns of topics: document topic density matrix θ and word topic density matrix ϕ. The word matrix ϕ has two dimensions K and V where K is the number of topics and V is the vocabulary size. Any value of $\phi_{k,v}$ represents the likelihood of word $v = 1, 2, ..., V$ belonging to topic $k = 1, 2, ..., K$. The document matrix θ has also two dimensions K and D where K is again the number of topics and D is the number of documents. A value of $\theta_{d,k}$ signifies the probability with which a topic $k = 1, 2, ..., K$ is likely to appear in a given document $d = 1, 2, ..., D$. Since LDA uses probability distributions from the Dirichlet family, it requires two Dirichlet priors; one for θ and another for ϕ. Each of the priors is governed by K (i.e. the number of topics) parameter and a prior parameter. It is referred to the prior parameter (i.e. α for ϕ and β for θ) as a model hyper-parameter which it affects the specificity of document-topic and word-topic distributions.

7.2.1.1.2 Non-Negative Matrix Factorization (NMF)

Unlike LDA, NMF is a deterministic algorithm that uses a decomposition technique for multivariate data where non-negative constraint is necessary for learning topics. It factorizes a high-dimensional data matrix $X = (X_{j,i})$ into lower-dimensional matrices A and B such that $X \approx AB$. The aim of the factorization is to find hidden themes (i.e. topics) within data. The values of X, A, and B and their coefficients are non negative. The X matrix is a term-document matrix with dimensions $D \times W$ where D is the number of documents and W is the number of words in the corpus vocabulary. $X_{j,i}$ represents the frequency of word j_{th} in document i_{th}. The frequency of words can be replaced by their corresponding TF-IDF weights. A is a document topic matrix with dimension $D \times K$ and B is a $K \times W$ word topic matrix, where K is the number of topics. A and B are computed by optimizing a loss function that is solved using gradient descent methods. Since we are dealing with large and unstructured datasets, we use the NMF algorithm developed by Renbo and Vincent [238]. The algorithm is optimized to handle the issues of processing large datasets and the existence of outliers.

7.2.1.2 Deep Learning Topic Modeling Approach

LDA and NMF topic modeling algorithms require efforts for hyperparameter tuning in order to predict meaningful clusters or topics. BERT-based (i.e. Transformer-based) topic modeling approach [86], on the other hand, alleviates this requirement by leveraging pre-trained language models (PLMs or LLMs) that learn the contextual representations of words unlike LDA and NMF that learn on count data and ignore the order and context of words. There are four main components of BERT-based topic modeling (BERTopic [86]):

- Transformer Embedding: this component is responsible for building dense embedding vectors by learning the semantic relationship between words using the transformer

architecture [86]. This step is very important to build a high quality embeddings as it will affect the learning of topic modeling and hence inferring comprehensive topics.

- Dimensionality Reduction: this component compressed the high dimensionality of the resulted embeddings into a lower-dimensional space. It is well-known that clustering algorithms are sensitive to the high dimensionailty of features [86]; hence this step is crucual for the clustering component to work efficiently. UMAP (Uniform Manifold Approximation and Production) [86] - a dimensionality reduction technique- is used to capture and preserve the global and local high-dimensional word representations (i.e. features) in lower dimensions.

- Clustering: this component is responsible for clustering the low-dimension embedding vectors into groups of similar embeddings. HDBSCAN [86] -a hierarchical density-based clustering technique- is used to cluster the resulted low-dimesnion embedding while handling irregular cluster shapes and identifying outliers at the same time. The advantage of HDBSCAN technique is that it does not force input data in clusters but recognizes that some data points could be outliers. This in turn helps to extract topic representations more accurately.

- Topic Extraction: this component extracts topics for the created clusters resulted in the clustering step. This step implements a class-based TF-IDF (c-TF-IDF) [86] -a modeified version of TF-IDF- technique to identify the most relevant key words given all data points in a cluster, independently of the clustering process.

7.2.2 Phrase Extraction

Phrase extraction is a process concerned with the automatic extraction of a set of representative phrases that express the aspects of textual contents [91]. Supervised and unsupervised methods have been widely used for phrase extraction [173]. In this work, we are interested in studying the unsupervised learning approach as it does not require annotated data. Manual data annotation for phrase extraction is prone to human subjectivity as well as it is inefficient; it not only takes a lot of time and requires a lot of effort, but it is also costly. Statistical and graph-based ranking approaches have been widely adopted to extract phrases from textual collections. TFIDF at n-gram [199] level is a well-known method used for statistical-based phrase extractions. However, one of its drawbacks is that it requires large data to produce good results. In addition, it needs to be combined with n-gram technique in order to process multi-word phrases, and this is computationally expensive and time consuming especially when using longer n-grams. Furthermore, n-gram considers n consecutive words but does not take into consideration the occurrences of words in a complete phrase or sentence. TextRank [141] and SingleRank [217] were among the first graph-based algorithms that were developed for phrase extraction. They use words co-occurrence information in order to find candidate phrases. Later on, SGRank [55] and PositionRank [75] algorithms proposed to incorporate statistical and positional information along with the information of words co-occurrences. These algorithms rely on natural language processing (NLP) techniques like POS and n-grams to form key

phrases. They utilize POS tagging to use lexical units of specific part of speech limited only to nouns [55,75,141,217], adjectives [55,75,141,217], or verbs [55]. Given the short expression and multilingual nature of social media data, this introduces two limitations: (1) ignoring important information residing in different lexical units other than nouns, verbs, and adjectives, (2) increasing the resource cost of having different POS tagging system for different languages. Another limitation of the previously mentioned algorithms is that they analyze words co-occurrences within a fixed sliding window. This disables the flexibility of fine-grained measurement for words associations within a collection of data (i.e. individual data subsets of topics). Rapid Automatic Keyword Extraction (RAKE) [181], a domain-independent language-independent algorithm for phrase extraction, is able to overcome the limitations in the previous studies. It undoubtedly fits the scope of this study for four reasons. First, RAKE overcomes the TFIDF limitation on small datasets inasmuch as it is designed to perform on dynamic-size individual documents (i.e. topic data subsets) rather than on the entire corpus. Second, it is not constraint to specific language structure; this greatly fits the unstructured nature of social media data that does not follow grammar conventions and is full of misspelled words. Third, it reduces the computational overhead of NLP tasks such as POS tagging and n-grams. Fourth, its flexibility allows it to extract phrases of all possible lengths, free of fixed-sliding-window constraint and without the additional computations of n-grams. This is beneficial in exploiting every possible piece of information within short messages which, in turn, will improve the quality of online social data interpretation.

7.2.2.1 Rapid Automatic Keyword Extraction (RAKE)

RAKE [181], a graph-based algorithm, was designed based on the assumption that a key word consists of multiple words that are rarely split by punctuation or stop words. Stop words could be canonical or uninformative words. The remaining words are assumed to be informative and are referred to as content words. RAKE takes two inputs: stop word list and punctuation list (i.e. word punctuation and phrase punctuation). The extraction process starts with splitting a given text into a set of candidate key words at the occurrence of pre-defined word delimiters. Next, the set of candidate keywords is split into a sequence of consecutive words at the occurrence of phrase delimiters and stop words. The consecutive words within a sequence together form a new candidate keyword (i.e. phrase). A graph of word-word co-occurrences is created to be used in computing the scores of the candidate words and phrases. Three scoring metrics were proposed: Word Degree $deg(word)$ calculates the words that have often occurrences in a document as well as in longer candidate phrases, Word Frequency $freq(word)$ computes the words that occur frequently without taking into consideration the word-word co-occurences, and Ratio of Degree to Frequency $\frac{deg(word)}{freq(word)}$. For the purpose of this work, we use the Degree of Words $deg(word)$ as a metric to compute phrases scores. The score of a candidate phrase is calculated as the sum of its words scores.

7.2.2.2 TextRank

TextRank [141] is a graph-based ranking algorithm used for keyword and phrase extraction. TextRank uses the word co-occurrence statistics to compute scores of words and extract phrases from texts. It uses the co-occurrence information to build a word graph. Two processes are applied before the graph is constructed: (1) words are filtered using POS mechanism and only nouns, verbs, and adjectives are used, (2) a sliding window value is defined. Each word in the graph represents a vertex and an edge between any two words is added if the two words co-occur within the pre-defined sliding window. A weight is assigned to every edge in the graph and its value represents the number of times a word co-occur within the sliding window. Each vertex is assigned with a score that reflects its importance and is computed in an iterative manner using PageRank algorithm. After convergence, the top n scored words are selected as keywords. Phrases are constructed if adjacent keywords are found in the resulted keywords.

7.3 Methodology

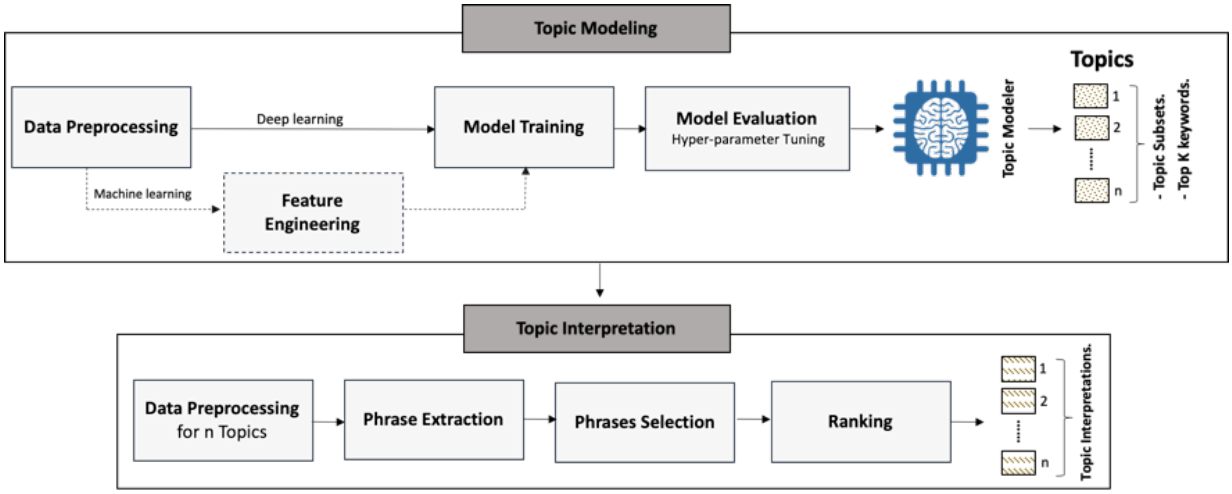


Figure 7.1: **Proposed methodology for data exploration and interpretation.**

The main objective of this section is to explore and find patterns in social media data and then generate explainable interpenetration of these patterns. Topic modeling is one approach to explore these patterns in large datasets and discover latent patterns (i.e. topic within data). The general framework used in unsupervised learning for topic modeling is followed in this thesis. In Figure 7.1 - Topic Modeling, the methodology used to build our topic model is illustrated. The data preprocessing of topic modeling is designed based on the criteria to increase the topic relevance and minimize uninformative parts of the data. According to the literature [67, 68], the data dominated by stopwords and general uninformative words is semantically uninterpretable as it reduces the reliability and utility of topic models. Accordingly, we consider removing two types of stopwords: (1) canonical

words ("the", "or") and dataset-specific words that have a very high and very low usage frequency. Vocabulary is limited to nouns, verbs, adjectives, and adverbs to increase the topic semantic coherence and to minimize the shortcomings of topic modeling algorithms like LDA and NMF, which treat all vocabulary words as having equal importance [137]. We adopt the suggestion made by Lau etal. [117] that lemmatizing data improves the topic coherence. The pre-processed data is then used to extract influencing features in the feature engineering component. Different types of features are investigated in the hope of finding effective ones that fit the short and unstructured nature of OSNs data (more details can be found in Section. 7.7). The topic model is trained and evaluated using the engineered features. To explore and learn the topics, unsupervised learning approach is used. During the model evaluation, a series of sensitivity tests of hyper-parameter tuning are run in order to find the optimal set of values that produce the highest semantic coherence across topics. The output of the topic model will be k topics. For each topic, we consider the top n keywords and the corresponding subset of data. It is important to mention that the traditional topic modeling approach is adopted for English data since the resources for preprocessing and feature engineering are quite adequate. This, in turn, allows to reduce the expensive computation and high-end hardware requirements by deep learning approach to perform the learning process.

Even though tradition topic learning approach is computationally lighter that deep learning approach, it requires sufficient hyperparameter tuning, prior topic size selection, and adequate resources and efforts to reduce data noises through preprocessing data, engineering and selecting high-quality features. Low-resourced languages like multidialectal Arabic lack such resources (e.g. POS, stemming, lemmatization, stopwrods, etc.) for each and every dialect. As a result, it is risky to implement traditional topic modeling methods on such noisy data like the multidialectal Arabic on social media. Contrary to traditional topic learning, deep learning is able to learn high and low level features directly from data without the involvement of domain experts [231]. In addition, deep topic learning like BERTopic [86] is less sensitive to data noises than traditional topic learning is. This is because BERTopic is able to locally and globally attend to important words and phrases even with the presence of some noise [86]. In this work, we adopt the deep learning approach to learn topics from multidialectal Arabic data.

After the themes (i.e. topics) have been inferred, they are fed into the topic interpretation component to facilitate the interpretations of the topics; hence providing us with a deeper understanding of the topics automatically. Using the top n keywords only is inefficient in interpreting the coherent meaning of topics [138]. A good interpretation of a topic should convey two characteristics: capturing the meaning of the topic, and distinguishing topics from one another. Single words fall short on these characteristics as they lack the context of phrases and sentences [138]. In more details, single keywords are too general and might miss the semantic relationship to form the main idea of the topics. Phrases, on the other hand, add context to single words, hence providing stronger coherence. Moreover, phrases by nature are broad so they are able to capture the overall meaning of topics [138]. In this thesis, we propose the use of unsupervised phrases extraction approach. Automatic Rapid Keywords Extraction (RAKE) algorithm is utilized to find topic phrases. Figure 7.1 - Topic Interpretation describes our proposed methodology to extract phrases of topics.

First, the data subset for each topic is pre-processed independently. This process is similar to that of topic modeling; however, in phrases extraction stopwords are not removed and all the part of speech tags are not pre-processed. Second, RAKE algorithm is used to extract keywords and phrases from each topic data subset. The weights of the extracted keywords and phrases are computed using the degree of word metric $deg(word)$ that calculates the words that have often occurrences in a document as well as in longer candidate phrases. Third, keywords and phrases based on the top n keywords (i.e. resulting from our topic model) are selected. Before selecting RAKE keywords and phrases, the keywords (i.e. resulting from our topic model) duplication across topics are removed. The reason to remove the duplication is to make unique interpretations that distinctively represent each topic. Finally, the keywords and phrases are ranked according to the weights of the corresponding keywords and phrase degrees. The output phrases have various lengths with minimum of two. To choose the optimum length of phrases, the average length of phrases for each topic is calculated. After calculating the average, phrases with more general dimension and phrases with more specific dimension than the average are considered. This is done by selecting shorter phrases and longer phrases than the average length.

7.4 Datasets

7.4.1 Topic Modeling Dataset

- **English Datasets**:

  Two English COVID-19 datasets were collected from Twitter using three filters: (1) geo-location coordinates for both USA and Canada, (2) English language only, and (3) Covid-19 related keywords such as "covid19, covid, corona, corona virus, virus, pandemic). One dataset was collected for USA and the other for Canada. COVID-19 related keywords were used to retrieve the data. The list of the keywords includes covid-19, covid19, covid, corona virus, corona, and virus. The collection was conducted in three periods over a duration that extends from December 2019 to November 2020: period-1: December 2019 to April 2020, period-2: May 2020 - August 2020, period-3: September 2020 - November 2020. We used geo-location coordinates to define the geographical regions to retrieve the tweets from. The two English COVID-19 datasets were used for modeling the unsupervised learning components of DEI and analyzig the COVID-19 pandemic in over the duration between December 2019 to November 2020 in both USA and Canada.

- **Multidialectal Arabic Datasets**:

  Two Arabic COVID-19 datasets were collected from Twitter during COVID-19 pandemic in 2020. Geo-coordinates for Lebanon and Saudi Arabia were used to collect COVID-19 data in Arabic language from both countries using Covid-19 related keywords, according to the owners of both datasets [5, 235]. The two Arabic COVID-19 datasets were used for modeling the unsupervised learning components of DEI and

analyzig the COVID-19 pandemic in two countries and two dialects: Lebanon (i.e. Levantine dialect) and Saudi Arabia (i.e. Gulf dialect).

7.4.2 Phrase Extraction Evaluation Dataset

Tsix dataset [157] was used to evaluate phrase extraction using RAKE algorithm on social media data (i.e. tweets). Tsix dataset consists of 32970 tweets that were categorized into six topics: brexit, election, isis, nobel, note7, and spacex. Each group of tweets (i.e. belonging to a topic) was assigned into a cluster. Each cluster was assigned a summary reference which composes of candidate sentences selected by two human annotators.

7.5 Data Preprocessing

preprocessing data is a very essential step in machine learning in general. It prepares the resource knowledge for the machine models to learn from. High quality preprocessing ensures the quality of the learning process. The objective of preprocessing data is to remove excess noise, which could affect the learning performance, and retain useful information. This work consists of two main components: data exploration and interpretation (DEI) and online social behavior modeling (OSB). Each component is processed independently as each requires a different preprocessing mechanism. Following are the preprocessing steps we propose to use:

1. **Removing HTML encoding symbols**
2. **Removing user mention**
3. **Removing URLs**
4. **Removing Retweets**
5. **Removing extra whitespaces**
6. **Converting text to lower case**
7. **Expanding abbreviations**: to replace abbreviation words with sequence-of-words format. This step processes contraction words "e.g. we're", negation words (e.g. "don't"), and slang words "e.g. ppl, bro". This is applied to English data only.
8. **Tokenizing text**
9. **Fixing repetition**: to remove character repetition and replace it with a single character. For example, word 'niiice' will be replaced by 'nice'. We believe in the importance of this step as it ensures the generality of learning.
10. **Converting iconic emotion into textual format**: to convert emojis and emoticons into textual representations [98].

11. **Removing special characters and numbers**

12. **Removing stop words**: We use two types of stop word lists: (1) standard list by standard libraries like NLTK, (3) customized list that is constructed manually by empirical experiments. According to the literature, this step has a major impact on topic predictions.

13. **Removing words with high and low frequencies**: to remove words of frequencies greater than 60% and less than 10 occurrences per the dataset.

14. **Tagging Part of Speech**: labeling words with grammatical description. We do this step to include only adjectives, adverbs, nouns, proper nouns, and verbs. The aim is to increase the efficiency of topic modeling performance. This is applied to English data only.

15. **Removing short words**: We assume that words with a single character does not have an independent meaning, hence, they do not contribute to the learning process. words of length less than 2 are removed.

16. **Lemmatization:** lemmatization to change each word into its original form. The objective is to reduce the size of vocabulary by conflating terms with related meaning. This is applied to English data only.

7.6 Experiment Design & Evaluation Protocol

The objective of topic inference (i.e. exploration) experiments is find comprehensive topics that represent given data in order to facilitate the analysis of online social behavior. We divide the experiments into two categories:

- **Traditional Topic Modeling**: The objective of this experiment category is to find a topic modeling algorithm and feature type that yield the best performance to find the optimal number of topics within given OSNs datasets. Two types of experiments were conducted for this purpose: (1) studying two topic modeling algorithms: LDA and NMF. (2) studying two types of features: BOW and TFIDF. We investigated the performance of the LDA and NMF with both features BOW and TFIDF on two datasets (i.e. COVID-19 collected during the period 1 - December 2019 to April 2020): (1) COVID-19 – Canada, (2) COVID-19 – USA. This yields a total of eight experiments for the topic modeling. We run a series of sensitivity tests to determine the best values of model hyper-parameters as summarized in Table 7.1. The tests were performed in sequential manner; one parameter at a time by keeping the others constant and then we run them over the datasets. For training, words with ≤ 10 occurrences and $> 60\%$ of occurrences in a dataset were filtered out. To assess the topic model performance in finding the optimal size, coherence score was used as an evaluation metric. Note than the experiments in this category are conducted on English data.

- **Deep Learning Topic Modeling**: BERTtopic [86] is used on multidialectal Arabic data to learn and infer representative topics from COVID-19 datasets for both Lebanon and Saudi Arabia. We use pre-trained Arabic language model as embeddings. Since we use a deep learning technique in this experiment, feature engineering and selection are not required. Also, it is not needed to define the number of topics in advance as BERTopic [86] uses HDBSCAN clustering algorithm that does not allow to specify the number of clusters in advance. We use two COVID-19 Arabic datasets to evaluate to Arabic dialects; one for Arabic-Levantine dialect [235], and another for Arabic-Gulf dialect [5]. We used coherence score [207] as a metric for our performance evaluation. A coherence score for a topic is calculated by measuring the degree of semantic similarity between high scored words within the topic.

Table 7.1: **The selected hyperparameters used for tuning before training LDA and NFM models.**

Model	Parameter	Value
LDA	Number of topics: K	Range from 2 to 14
	Dirichlet hyperparameter: alpha α	Range between 0.01 to 1.
	Dirichlet hyperparameter: beta β	Range between 0.01 to 1.
NFM	Gradient Descent step size: kappa	Range between 0.1-0.5
	Number of topics: K	Range from 2 to 14

For topic interpretation (i.e. phrase extraction), we found out that the phrase length of three has the highest average frequencies in all the topics for Canada, USA, Lebanon, and Saudi Arabia datasets. We considered the lengths of two and four to add a more general dimension before and a more specific dimension after, than the phrases of length three. We evaluated RAKE algorithm on Tsix dataset using ROUGE metric [126]. ROUGE, Recall-Oriented Understudy for Gisting Evaluation, is a well-known metric used for evaluating automatic summarization of texts. It works by comparing the automatically generated summaries to human-generated summaries. ROUGE-N metric computes the overlap of n-grams between the automatic and human summaries. In this work, we are interested in evaluating the recall of our phrase extractor by examining the percentage of n-grams, in our generated phrases, that exist in the reference phrases. In addition, we compare RAKE algorithm to other two phrase extraction algorithms, TextRank and TFIDF using the same dataset. In this thesis, we set the n-gram to 4-gram and therefore the evaluation was performed using ROUGE-1, ROUGE-2, ROUGE-3, and ROUGE-4. In this thesis, we use keywords (i.e. single words) and phrases of lengths 2-4 and hence the decision to use 4-grams in this experiments. Accordingly, TFIDF and TextRank models were built at the 4-gram level. We were able to fix the sliding window for TextRank to maximum 3 due to hardware limitation.

7.7 Results and Analysis

7.7.1 Topic Modeling

We present the results of the methodology we follow for modeling key topics using OSNs short texts. For this experiment, we use two datasets related to COVID-19; one for Canada and another for USA. It is worth mentioning that stopword filtering has shown a major impact on the overall topic modeling learning. The following experiments were conducted on all the data after removing stopwords.

7.7.1.1 Traditional Topic Modeling Approach

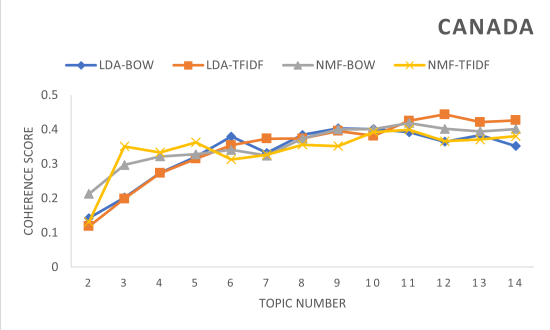
7.7.1.1.1 Learning Features

To find the best representative features, we have considered using three types of textual features: Bag-of-Words (BOW), Time-Frequency-Inverse-Document-Frequency (TFIDF) and n-gram. Our empirical experiments have shown an improvement in learning when *uni*-gram and *bi*-gram were combined together. To find the optimal number of topics, two experiments were conducted by combining uni-bi-gram with BOW and TFIDF. Figure7.2 illustrates the performance results of LDA and NMF models using BOW and TFIDF in order to find the best topic model for OSNs data. The BOW and TFIDF features were constructed based on uni-bi-gram features. According to the results, LDA model performed better with TFIDF features than it did with BOW features in both datasets. Unlike LDA model, NMF model performed better with BOW than it did with TFIDF.

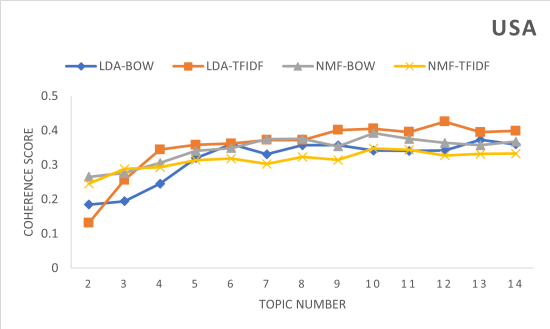
Overall, LDA model trained using TFIDF features outperformed NMF model trained using BOW features. For Canada dataset, LDA-TFIDF model maintained the highest scores for topic sizes of 11 to 14. It also scored the highest for topic size of 7. Similarly, LDA-TFIDF model maintained the highest scores for topic sizes of 4 to 14 in USA dataset. This shows that TFIDF method works well with OSNs texts since they are huge in volume while short in length and hence the content is limited per message post. As a result, the need to spot influencing words, to learn different topics, arises. TFIDF method, which plays at word level, measures the relevance of words but not the frequency; it represents documents about "computers", for example, far from documents about "batteries". This gives it the advantage of choosing influencing vocabulary and reducing the complexity of training since using the entire vocabulary in training [36] is expensive. By using the TFIDF weights, the chance that rare words are sampled would increase (i.e. which is the goal to improve the topic learning on short texts documents of a large size). This results in making them have a stronger influence on topic assignment. This is the same reason why it is recommended to remove stop words before training an LDA model.

7.7.1.1.2 Topics Size

Table 7.3, shows the results of the hyper-parameter tuning for the LDA and NMF models. The best values of the model hyper-parameters have been chosen based on the top 10



(a) **Canada**



(b) **USA**

Figure 7.2: **Performance of four topic models to find the optimal topic size for Canada and USA, in terms of coherence score.**

highest coherence scores. The set of hyper-parameters values that has the majority of the highest coherence score are selected. The values of LDA alpha α and beta β and NMF kappa are shown to be fixed for the top 10 highest coherence scores. For the number of topics parameter K, it ranges between values of 6 and 14. However, no exact value was given. Therefore, we selected the best hyper-parameters values as suggested in Table 7.3 to determine the exact number of topics for topic modeling training. Note that we select 2 as the minimum number of topics and 14 as maximum. We then run another process of tuning with respect to the number of topics. Figure7.2 illustrates the performance of LDA and NMF models over a range between 2 and 14 topics with the hyper parameters alpha α and beta β fixed at 0.01 and 0.51 for LDA models, and kappa fixed at 0.2 and 0.3 for NMF-BOW and NMF-TFIDF models, respectively. Twelve topics are shown to score the highest coherence by LDA-TFIDF model for both datasets before the performance decreases and flattens out. Similar observation was claimed by Yuxin et al. [51] where LDA showed more robust performance on full sentences than NMF did. To find the optimal number of topics for the other two periods (i.e. May 2020 - August 2020, September 2020 - November 2020), we follow the same steps mentioned previously using LDA-TFIDF. Table 7.2 summarizes the optimal number of topics resulting from LDA-TFIDF models for Canada and USA datasets. The results represents three periods during the pandemic: December 2019 - April 2020, May 2020 - August 2020, September 2020 - November 2020).

Table 7.4 shows the results of our LDA-TFIDF model on Canada and USA datasets. We select the top 30 keywords in this thesis but due to a limited writing space, we only

Table 7.2: **The optimal number of topics for Canada and USA datasets for three periods during the pandemic. The optimal topic sizes were determined based on the highest coherence scores for each dataset.**

	Canada		**USA**	
	Topic Size	**Coher-Score**	**Topic Size**	**Coher-Score**
Period-1	12	0.44	12	0.43
Period-2	13	0.4	10	0.4
Period-3	10	0.43	11	0.5

list the top 20 keywords of sample topics inferred by our LDA-TFIDF model during for three periods during the pandemic. Note that the keywords listed in the table are before removing the duplications across topics.

Table 7.3: **The optimal hyperparameters values for LDA and NFM models, resulted from hyper-parameter tuning process using two COVID-19 datasets.**

Model	Parameter	Best Value
LDA	Number of topics: K	6-14
	Dirichlet hyperparameter: alpha α	**0.01 for both BOW and TFIDF**
	Dirichlet hyperparameter: beta β	**0.51 for both BOW and TFIDF**
NFM	Gradient Descent step size: kappa	**0.2 for BOW**, **0.3 for TFIDF**
	Number of topics: K	7-14

Table 7.4: **Top 20 keywords for sample topics inferred by LDA-TFIDF model, for Canada and USA datasets, during three periods of the pandemic.**

	Canada	**USA**
	Period 1	
Topic 1	love, quarantine, life, hope, home, stayhome, pandemic, walk, safe, dog, work, fun, hair, isolation, toronto, enjoy, song, kid, ontario, normal, friend	quarantine, nyc, lockdown, song, york, stayhome, pandemic, quarantinelife, love, youtube, stayathome, brooklyn, movie, home, street, walk, elon_musk, photo, fun, staysafe
Topic 2	trump, china, americans, chinese, president, world, response, america, blame, bad, death, die, country, lie, government, kill, pandemic, vaccine, fauci, believe, white_house	trump, china, americans, president, obama, lie, blame, response, death, chinese, die, election, country, america, pandemic, world, hoax, government, state, kill
Topic 3	canada, news, cure, world, petition, big_banks, canadian_authoritie, publish, media, expert, chronic_ink, elon_musk, poor, investigation_stock, fight, warrant_senator, superbug, contact_trace, conspiracy_theorie, cargill_close	spread, sick, die, mask, stop, home, bad, work, wear_mask, stupid, safe, kid, open, wear, test, life, wearing_mask, understand, risk, crazy
	Period 2	
Topic 1	case, death, test, mask, report, health, number, die, spread, vaccine, positive, ontario, canada, school, flu, wear, cdc, testing, risk, news, patient, confirm	pandemic, vaccine, test, health, death, patient, work, school, case, risk, impact, cdc, testing, learn, die, business, crisis, spread, report, care
Topic 2	hair, beautiful, pretty, wear, idk, mask, cut, hurt, eye, face, head, black, white, love, brain, nose, ask, point, cough, blue, lady, cat	nan, love, eat, bad, food, taste, chicken, thread, pizza, dinner, karen, cook, slogan, cake, meat, ready, boo, lunch, yup, ugh
Topic 3	trump, die, china, americans, death, president, kill, country, lie, america, dead, bad, vote, stop, world, response, state, blame, pandemic, hoax, election, biden	trump, death, die, americans, china, kill, president, country, dead, america, lie, biden, spread, vote, pandemic, response, state, rnc, mask, stop
	Period 3	
Topic 1	trump, biden, election, vote, president, americans, die, stop, joe, lose, country, america, lie, win, disappear, believe, plan, death, rally, task_force, president_elect, hoax	trump, die, americans, biden, country, kill, lose, dead, president, lie, care, american, china, vote, economy, life, america, election, job, death
Topic 2	pandemic, business, impact, work, support, health, job, community, learn, pay, challenge, market, canada, program, money, plan, care, crisis, service, recovery, economy, government	mask, wear, spread, celebrate, rally, trump, stop, super_spreader, event, biden, party, crowd, celebration, safe, social_distancing, street, social_distance, forget, supporter, sick
Topic 3	vaccine, pfizer, effective, news, early, trial, pfizer_biontech, cure, study, test, mink, denmark, science, announce, infection, mutate, human, antibody, research, candidate, develop, data_signal	vaccine, pfizer, test, positive, effective, ben_carson, news, election, disappear, pfizer_biontech, chief_staff, mark_meadow, trump, trial, early, result, white_house, announce, percent_effective, analysis

7.7.1.2 Deep Learning Topic Modeling Approach

Table 7.5: **The performance of BERTopic models in term of coherence scores on Lebanon and Saudi Arabia COVID-19 datasets. The number of topics is determined automatically by BERTopic during training.**

	Topic Size	**Coher-Score**
Lebanon	100	0.1
Saudi Arabia	100	0.1

Table 7.5 illustrates the performance of two BERTopic-based models that have been trained on two Arabic datasets; COVID-19 in Lebanon (i.e. Arabic-Levantine dialect) and COVID-19 in Saudi Arabia (i.e. Arabic-Gulf dialect). Both are shown to perform at 0.1 of coherence score on topic size of 100 that has been determined by BERTopic during the training process. It can be observed from Table 7.3 that the inferred topics are meaningful and that the top terms for each topic provide a general idea about the inferred topics. For instance, topic 1 in Lebanon talks about Jordan during the COVID-19 pandemic whearas topic 1 in Saudi Arabia talks about staying home and quarantine within Saudi Arabia. COVID cases reports and death tolls are seem to be discussed in topic 9 in Lebanon and topic 2 in Saudi Arabia. Topics 6, 7, and 8 -in Lebanon- discuss political issues that were ongoing during COVID-19 pandemic in 2020. Topics 8, 9, and 10 - in Saudi Arabia- discuss local issues that happened during the COVID-19 pandemic back in March 2020. Saudis are seem to have had positive times during the COVID-19 quarantine and that is shown in the activities topic inferred from the Saudi dataset. Both Lebanon and Saudi Arabia have communicated soccer sport during the COVID-19 pandemic in 2020 and that is clear in topic 18, 17 in Lebanon, Saudi respectively.

Topic	Lebanon	Saudi Arabia
Topic 1	اردن, تجول, حكومه, اربد, حظر	خليك_بالبيت, الحجرالمنزلي_واجب_وطني, خليك_في_البيت, قاعد_بالبيت, حظر_تجول
Topic 2	اليمن, عدن, حوثي, يمني, يمنيه	تسجيل, جديدة, إصابة, حالة, متحدث, بفيروس
Topic 3	ايران, ايراني, ايرانيه, كرونا, طهران	المعقمات, بلاغ, الاسعار, الكمامات, الصيدليات
Topic 4	قطر, دوحه, قطريه, حمدين,, تميم	اليدين, غسل, بالماء, زيبارت, والصابون
Topic 5	الاسد, سوريا, سوري, ادلب, بشار	العراق, عيب, كورونا_العراق, العراق_ينتفض, الوقاية
Topic 6	ايغور, مسلمين, الصين, مسلم, ربك	اليمن, العدوان, اليمني, يمني, الحوثي
Topic 7	احتلال, غزه, اسرائيلي, اسرائيل, اسرى	الامارات, مواطني, ملتزمون_ياوطن, الإمارات, بوخالد_الغذاء_والدواء_خط_أحمر
Topic 8	صفقه, صهاينه, العرب, جزيره, شيطان	الخاص, القطاع, العمل, الموظفين, الشركات
Topic 9	حاله, وفاه, اجمالي, مصابين, وفيات	الطلاب, التعليم, الطالبات, الطالب, تدريب_الامام_كليه_اللغات
Topic 10	حظر, منزل, بيتك, تجمعات, الحظر	تعليق_الصلاه, الصلاة_في_رحالكم, يوم_الجمعة, الجمعه, صلوا_في_رحالكم
Topic 11	ليمون, تشرب, العسل, اعراض, ركز	فرجت_مع_الراوي_دخيل, فاتوره, المبلغ, الخير, فزعتكم
Topic 12	يدين, تجنب, غسل, لمس, سعال	بلغنا, رمضان, رفعت, فاقدين, مفقودين
Topic 13	الخاص, قطاع, توظيف, وفصلوني, لفصلي	شكرا, أبطال_الصحة, أبطال, أعماق, لأبطال
Topic 14	الشعب, فاسدون, نقاش, مشاركه, دعم	الكهرباء, الفواتير, ماء, المشكله, المياه
Topic 15	مؤشر, نقطه, السوق, مخاوف, دولارا	مسلسل, مسلسلات, نتفلكس, فيلم, افلام
Topic 16	النفط, اسعار, نفط, بترول, دولارات	وضع, الامهات, جنة, المتزوجين, الان
Topic 17	طلبيه, الشحن, ملابس, اكسبرس, ارمكس	النصر, الهلال, الدوري, ايقاف_الدوري, حمدالله
Topic 18	برشلونه, مدريد, ريال, ميسي, بهدل	كود, خصم, نون, الخصم, كوبون

Figure 7.3: **Top 5 keywords for sample topics inferred by BERTopic-based models, for Lebanon and Saudi Arabia COVID-19 datasets.**

7.7.2 Topic Interpretation

Table 7.6 shows the performance of phrase extraction models using a tweet dataset (i.e. TSix). The evaluation was conducted on the top 30 phrases resulted from each model. At 4-gram level, RAKE model has been shown to outperform TFIDF and TEXTrank models in terms of execution time and ROUGE-n recall scores. RAKE model was able to recall 80% of the single keywords existed in reference sentences whereas TextRank and TFIDF recalled 60% and 50% of the keywords presented in the reference sentences. At 2-gram level, RAKE model was the winner in recalling 60% of phrases with length 2, followed by TextRank model with 40% of length-2 phrases recall and 25% recall by TFIDF model. Similarly with phrases of lenth 3 and 4, RAKE model yielded the best performance followed by TextRank and TFIDF models. Even though TextRank was shown to be the second best, it required $\approx$ 12 more times (i.e. 972 seconds) than RAKE (81 seconds) did. Other limitations of TextRank are fixed sliding window and phrase length boundaries; it is not able to dynamically extract phrases of varying lengths. This is shown in its ability to recall single words (60% of recall) while its performance deteriorated for phrase extraction. Like TextRank, TFIDF model also suffer from the limitation of phrase length boundaries; n-gram has to be fixed prior phrase extraction. In addition, TFIDF model has shown to have a weak performance in extracting keywords and phrases from small-sized data. The average data size per a cluster (i.e. document) in TSix dataset is 36 tweets. This is not surprising as TFIDF algorithm needs a decent amount of data in order to find influencing

words with respect to a whole document.

These results show the effectiveness of RAKE model for phrase extraction on OSNs data. Its advantages lay in its dynamic ability to extract phrases of varying lengths (i.e. not constrained to fixed sliding window nor fixed phrase length) and it works at a message level (i.e. it is not constrained to either small nor large datasets).

Table 7.6: **A comparison between RAKE, TFIDF and TextRank algorithms for phrase extraction using Tweet TSix dataset. The performance is evaluated in terms of execution time and ROUGE-n recall metric; where** $4 \geq n \geq 1$

	TFIDF	**TextRank**	**RAKE**
ROUGE-1	0.5	0.6	**0.81**
ROUGE-2	0.25	0.41	**0.6**
ROUGE-3	0.14	0.31	**0.44**
ROUGE-4	0.06	0.21	**0.32**
Execution Time (seconds)	82	972	**81**

Table 7.7 and Table 7.8 list the results of our RAKE-based phrases extracted from our COVID-19 datasets of Canada and USA. The tables show a sample of topics keywords and phrases for Canada and USA during the three periods defined in this thesis. The keywords in the Tables are the output of our LDA-TFIDF model after removing the duplicates across the topics and then ranking them based on their RAKE weights. We can detect an improvement in the quality of the keywords after removing the duplicates. For example, in Canada topic 3 - Period-1 (Table 7.4), we see that the top 4 keywords are common in other topics. This duplication not only degrades the quality of the topic description but also adds ambiguity. After removing the duplications we notice that the comprehensibility of the topic has enhanced (Table 7.7). RAKE produces scores based on the frequency of phrases in each topic. Therefore, the high frequency of phrases of a certain topic (i.e. after choosing the most influencing words resulted from the LDA-TFIDF model) indicates that people are indeed talking about this topic. Thus, it has become clear that the topic 3 discusses Elon Musk and Tesla. Additionally, we see that the inferred phrases (Table 7.7) take it further and show that topic 3 is about Elon Musk and Tesla in the Canadian news (e.g. "social media bbc news", "cure cbs news") and that Elon Musk might reflect negative vibes (e.g. "elon musk threaten", "fight elon musk tweet") during the pandemic in Canada.

Looking at the keywords of topic 4 in Canada (Table 7.7), we are unable to understand what the main idea of the topic is. They only provide names of prominent figures like British Prime Minister "Boris Johnson", American government official "Stephen Miller", and public events like "UFC" (Ultimate Fighting Championship). Thanks to the topic interpreter that has provided details about the keywords by adding contexts, which facilitated the understanding of the topic. For instance, phrases of length 3 have added some information about Boris Johnson; the phrases "boris johnson die" and "save boris johnson" bring forth the idea that Boris Johnson is undergoing some critical situation. The

idea of the topic is wrapped up in phrases of length 4; now we know that Boris Johnson has been tested positive "boris johnson tests positive", and this explain the phrases "boris johnson die" and "save boris johnson". In the case of Stephen Miller, however, we are able to conclude that he was tested positive "stephen miller test positive" in a phrase of length 4 only while no information was mentioned in phrases of lengths two and three (i.e. top 10 phrases of length 2 and 3). With respect to UFC, phrases of length 2 are about cancelling ufc: "ufc cancel". Phrases of length 3 add more details to phrases of length 2; we now know that the fight game cancellation is associated with the player Jacare, a Brazilian mixed martial artist,: "fightcancelled jacare ufc". Also UFC has something to do with testing positive "ufc card test positive" (i.e. a phrase of length 4). We already have the knowledge, from phrases of length 3, that UFC game cancellation might be associated with the player Jacare. Put together we can infer that Jacare might have been tested positive and that is why the game was cancelled. Another example can be seen in USA Topic 4 - Period-1 (Table 7.8). Phrases of lengths 2, 3 and 4 add more dimension to the topic readability and comprehensibility than single keywords do. In phrases of length 2, we notice that sport season is the topic talked about the most. Phrases of length 3 explain further and mention spring sport season, high school sport, sports season during pandemic and NFL/NBA season, for instance. They reveal also that the topic is about travel season and future travel plans. Phrases of length 4 add up more details to the season, sport and travel that they were cancelled. Additional information was also revealed, in phrases of length 4, that baseball was among the sport activities in high schools. It is worth noting that such details are not highlighted when using LDA's single keywords exclusively.

Table 7.4 lists the results of our RAKE-based phrases extracted from COVID-19 datasets of Lebanon and Saudi Arabia. The table shows a sample of topic keywords and phrases used during COVID-19 pandemic over there. By looking at the keywords of topic 3-Saudi Arabia (Table 7.4), we can see that these keywords are mainly about the private sector: "القطاع الخاص" meaning private sector, "العمل" work, "القطاع" sector, "الموظغين" employees, "الشركات" companies, "الحكومي" governmental, and "موظف" an employee. However, this set of single keywords is not contextualized, therefore, the message is not conveyed. Thanks to our topic interpreter that provides details about the keywords by adding contexts which has facilitated the understanding of the topic. Phrases of length 3 and 4 have added some information about the private sector and employees; the phrases of length 3 ("القطاع الخاص ضد" meaning private sector is against, "موظفين القطاع الخاص" private sector's male employees, "موظفات القطاع الخاص" private sector's female employees) indicate that the private sector is against its employees of both genders male and female. The idea of the topic is wrapped up in phrases of length 4; the phrases
"القطاع الخاص الصحي مظلوم" meaning the private health sector is oppressed,
"القطاع الخاص يتجاهلون القرارات" private sector is ignoring the decisions,
"اغلاق القطاع الخاص مطلب" shutting down the private sector is a requirement,
"إلزام القطاع ـ الخاص بتعليق العمل" forcing the private sector to suspend work, have added more context to complete the idea of the topic; the employees of private sectors seem to have been complaining about the private sector not abiding by the Corona virus measures im-

plemented by the authorities during the early stages of the pandemic back in 2020, and demanding that the employees go to the work place instead of quarantining at home. Similar to topic 3-Saudi Arabia, the phrases of length 2, 3 and 4 ("مسلسلات نتفلكس" meaning Netflix series, "مسلسل جديد" a new tv series, "مسلسل فاست" the Fast tv series, "مسلسل تركي" a Turkish series, "فلم افلام نتفلكس" Netflix movies, "ومافيه زي مسلسلات زمان" nothing like old tv series, "جبت مسلسلات بالنتفلكس مظلومه" I found Netflix series that are underrated, "نصائح ل مشاهدات افلام" recommendations of movies to watch, "فلم الليله جميل فعاليات" the movie tonight nice activities), have wrapped up the context of topic 8-Saudi Arabia, whose single keywords set includes "مسلسلات" meaning tv series, "فلم" a movie, "طاش" Tash is an 80's Saudi comedy show, and "مشاهدات" views. Unlike the single keywords that have failed to convey the message, those phrases have manifested that some people in Saudi have been watching globalized movies and series, giving recommendations, and voicing out their opinions on movies. Further, we are unable to understand what the main idea of topic 9 in Saudi Arabia is, by just looking at the keywords. The set of single words that includes ("وضع" meaning a situation, "جنة" heaven "الامهات" mothers, "زلمة" a guy, "المتزوجين" married couples, "الزوجات" the wives) does not communicate a clear message. In contrast, phrases of length 2 bring forth that the topic talks about the back then situation of mothers and married couples at home during the pandemic. New variables, children and fathers, have been detected in phrases of length 3 ("وضع الآباء الان" meaning fathers's situations now , "حال الامهات اطفالهم" mother's situation their children). Those variables have provided new details that serve the over all message of the topic. The other longer phrases like "افرقنا زلمة معاناة الزوجات" meaning get lost man! wives' sufferings" and "الأنثى الزوجة تجعل المكان جنة جحيم" the female spouse makes the place heaven or hell, conclude that those mothers and fathers have been suffering during the quarantine. The latter phrase reflects a universal cultural concept that female spouses are responsible for the happiness or the misery of marriage life, "happy wife happy life".

Phrases of maximum lengths are mostly discarded since long phrases (i.e. sentences) do not provide the overall meaning of the topics. instead they only show one part of the whole topic. A case example can be seen in Canada topic 4 - Period-1 (Table 7.7). The maximum-length phrase says "prime minister british prime minister boris johnson tests positive". The long phrase only shows one part of the topic and did not show the whole picture that the topic was talking about famous figures who tested positive. Another example can be seen in topic 8 of Saudi Arabia (Table 7.4). The long sentence
"اكملت للتو فلم للمخرج مارتن سكورسيزي والمصور روبيرت ريتشاردسون امسيت ثملا بسحر الفلم متنهى الغموض النهاية متوقعة ابدا أيهما الأفضل"
meaning "I just finished watching a movie directed by Martin Scorsese and filmed by Robert Richardson. I was enchanted by the thrilling and unpredictable aspect of the movie...", provides a specific sub-detail about the topic which, again, shows only one part of the whole topic. This finding is supported by the results obtained by Qiaozhu [138] that sentences

might not be accurate to capture the general meaning of a topic as they might be too specific.

Overall, the results show the effectiveness of our topic models and phrase extractors in automatically identifying and interpreting topics inferred from OSNs data. It has big benefits in minimizing the human intervention in identifying and interpreting topics which in turns facilitates the real-time topic modeling and interpretation with minimized need to human approvals.

The main limitation of this component is the non-dynamic topic modeling that does not analyze the evolution of topics over time. Despite this limitation, our proposed topic models have shown good performance results in discovering patterns and inferring topics from the challenging noisy unstructured-formatted OSNs data. For traditional topic modeling, combining TFIDF with NLP techniques and carefully preparing our data and crafting our features have successfully contributed in building topic models that are capable of handling the OSNs data (i.e. English in this work), as reported in our results and analysis. Deep learning approach has reliably solved the problem of insufficient NLP preprocessing tools for Arabic multi dialects; BERTopic model has shown a superior performance in inferring meaningful topics from OSN dialectal-Arabic data while maintaining robustness against the noise residing within the data.

Table 7.7: **Top keywords and phrases extracted based on LDA-TFIDF top keywords and Ranked based on RAKE - Canada.**

		Keywords	Phrases-L2	Phrases-L3	Phrases-L4	Phrases-Lmax
Period-1	Topic 1	stayhome, quarantine, isolation, staysafe, stayathome, quarantinelife, beautiful, dog, fun, walk, song, anxiety, hair	stayhome quarantine, quarantine stayhome, socialdistance stayhome, stayhome socialdistancing, canada stayhome, stayhome canada, stayhome toronto, home stayhome, toronto stayhome	socialdistance quarantine stayhome, stayhome quarantine pandemic, stayhome staysafe quarantine, stayhome staysafe socialdistancing, staysafe stayhome socialdistancing, toronto stayhome ontario, home staysafe stayhome, stayhome socialdistance flattenthecurve, pandemic stayhome toronto, stayhome stayhome selfquarantine,	quarantine quarantine stayhome staysafe, quarantine social distancing stayhome, pandemic mentalhealth quarantine stayhome, stayhome isolation quarantine ottawa, untiltomorrow stayhome quarantine ontario, pandemic stayhomesavelive stayhome socialdistancing, quarantinelife toronto staysafe stayhome, stayhome staysafe socialdistance besafe, stayathome stayhome socialdistancing washyourhand, distance socialdistance stayhome plankthecurve	canada health government protectyourself stayhome socialdistance life lifelesson staysafe montreal toronto ottawa canada usa nyc california europe greece
	Topic 2	white_house, president, chinese, america, americans, american, fauci, obama, racist, blame, believe, election, bad	president trump, trump americans, trump chinese, trump president, chinese trump, trump america, lie trump, americans trump, america trump, obama trump,	china trump chinese, trump blame china, white house trump, government president trump, president trump chinese, break president trump, updates president trump, america president trump, trump wartime president, stop president trump,	trump americans trump trump, trump president trump shame, trump america save trump, trump trumpvirus trump bad, raise china president trump, trump stop blame china, trump blame china plan, donald trump china blame, buddy trump blame china, president donald trump miracle,	trump americans trump trump
	Topic 3	elon_musk, tesla, cure, contact_trace, true, future, poor, expert, petition, chronic_ink, infectious, publish, superbug	elon musk trump true, true canada, canada expert, expert world, trump cure, tesla cure, official tesla, tesla elonmusk, paraphrased tesla,	tesla elon musk, elon musk tesla, cure cbs news, elon musk threaten, random elon musk, elon musk reopen, elon musk furious, elon musk defy, elon musk lash, reopens tesla factory,	petition government canada canada, elon musk tesla factory, elon musk reopen tesla, cure minutes cbs news, social media bbc news, elon musk threaten move, fight elon musk tweet, elon musk champion defy, trump cure injecting disinfectant, bloomberg elon musk lash,	publish canada news media news
	Topic 4	boris_johnson, ufc, stephen_miller, murder_hornets, fraud, national_exhibition, whistleblower, scam, question, joke, meme, covidiot, answer	ufc jacare, boris johnson , cancel ufc ufc card, card ufc, fbi ufc, ufc cancel, government whistleblower, ufc fighter,	canadians question trump, fight cancel ufc, fightcancelled jacare ufc, murder hornets canada, save boris johnson, boris johnson die, dana white ufc, fight later ufc, boris johnson advice, boris johnson message,	ufc card ufc ufc, boris johnson tests positive, ufc card test positive, stephen miller test positive, contracts ufc ufc release, ufc card jacare souza, semen trump stephen miller, british boris johnson lay, whistleblower warn darkest winter, bank canada ctv news,	prime minister british prime minister boris johnson tests positive
Period-2	Topic 1	case, public, testing, ontario, news, risk, vaccine, hospital, number, high, patient, report, flu, cdc, infection, confirm, symptom	public health, death case, case death, case high, case birx, ontario health, case study, canada case, infection health	hit case case, toronto public health, case flu case, health case recover, case health cbc, additional case health, henry public health, ontario public health, case health unit, total case case	case death case death, prison health public health, case public health unit, positive case positive case, case deaths total death, case death active case, total number case health, case death total case, case report department health	case health authority vancouver coastal health fraser health island health interior health northern health
	Topic 2	hair, black, white, face, beautiful, blue, pretty, body, head, woman, cut, hot, cat, eye, brain, perfect, nose, lady, pain, idk, imagine, point, hurt, ask, haircut, cough	hair love, black hair, care hair, white hair, white love, hair mask, hair face, cut hair, hair blue, hair feeling	black hair salon, black woman hair, white hair head, love short hair, yoongi black hair, love hair colour, black hair dye, naturally black hair, cut hair love, love black woman	white black face mask, pink hair blonde hair, cut hair cute love, curler hair hair thick, black white green red, swing blue radio love, mask white black combo, friend dye hair red, blame love silver hair, personality hair white tooth	black trans woman black woman white gay black woman simple math
	Topic 3	trump, president, china, america, biden, americans, government, stop, dead, american, world, vote, country, response, economy, election, lie, control, kill, blame, protest, hoax	president trump, trump china, trump americans, trump president, trump america, biden trump, trump stop, job trump, trump world	trump anti trump, trump trump imbecile, donald trump trump, trump election trump, understand trump trump, trump trump republicans, trump blame trump, trump hoax trump, trump supporter trump, trump riot trump	trump blame china trump, elect president trump trump, america trump presidency trump, donald trump donald trump, trump administration donald trump, trump bad trump responsible, ivanka trump trump warrior, trump family trump hiding, trump kill nonsense trump, trump divide trump failure	trump killer murderer trump killer murderer trump killer murderer trump killer murderer trump killer murderer trump killer murderer trump killer murderer
Period-3	Topic 1	biden, president_elect, president, vote, americans, mail, america, joe, task_force, china, dead, american, rally, country, lie, believe, hoax, disappear	biden trump, trump biden, president trump, vote trump, trump president, americans trump, trump vote, trump americans, mail trump, american trump	trump rally trump, trump lie trump, trump trump lie, election trump biden, biden trump election, biden president trump, trump president biden, trump biden president, biden vote trump, trump biden vote	trump biden concede trump, trump biden blame trump, vote trump trump negligence, trump gop rally trump, trump rally trump rally, trump biden election election, impressive election trump biden, usa vote trump biden, compare president trump biden, die vote trump biden	president trump black matter trump trump
	Topic 2	support, community, market, economic, online, global, mental, canadian, recovery, service, impact, program, learn, economy, industry, crisis, job, pay, money, challenge, fund, company	mental health, community health, health online, health economic, economic health, online health, global health, health program, health crisis, pandemic support	pandemic mental health, mental health support, health care community, support mental health, health care service, health care job, canadian public health, public health service, work mental health, global health crisis	health care mental health, health issue mental health, financially health mental health, strengthen health care support, health care social service, support child mental health, pandemic affect mental health, economy health care education, relief health care job, support employee mental health	believe science believe health care professional believe alternative care professional believe mental health care expert
	Topic 3	vaccine, pfizer_biontech, pfizer, effective, study, science, mink, early, denmark, trial, research, human, fauci, drug, treatment, antibody, scientist, develop, candidate, cure, announce, mutate, prevent	study vaccine, pfizer vaccine, vaccine fauci, vaccine pfizer, vaccine news, vaccine trial, news vaccine, vaccine effective, vaccine trump, vaccine shot	vaccine quickly vaccine, vaccine secondly vaccine, vaccine pfizer drug, vaccine pfizer pfizer, vaccine news pfizer, pfizer vaccine news, pfizer vaccine effective, vaccine pfizer effective, effective pfizer vaccine, effective trump vaccine	pfizer vaccine mrna vaccine, vaccine effective vaccine brampton north, pfizer early vaccine data, vaccine vaccine merely protection, vaccine candidate effective pfizer, pfizer early data vaccine, amazing news pfizer vaccine, vaccine early trial news, vaccine news cnbc news, pfizer vaccine effective preventing	vaccine early data pfizer vaccine

Table 7.8: **Top keywords and phrases extracted using RAKE based on LDA-TFIDF top keywords. The keywords and phrases are ranked based on RAKE - USA.**

		Keywords	Phrases-L2	Phrases-L3	Phrases-L4	Phrases-Lmax
Period-1	Topic 1	nyc, quarantine, stayhome, brooklyn, staysafe, quarantinelife, manhattan, stayathome, elon_musk, street, beautiful, song, youtube, murder_hornets, fun, walk, dance, movie, photo, enjoy, wearamask, tesla, laugh	quarantine nyc,nyc quarantine, nyc stayhome, stayhome nyc, stayhome quarantine, quarantine stayhome, socialdistance nyc, nyc socialdistance	quarantine stayhome quarantine,nyc socialdistance stayhome,nyc lockdown nyc, quarantine quarantinelife nyc, quarantinelife quarantine nyc, stayhome quarantinelife nyc, stayhome quarantine pandemic, nyc nyc nyclockdown, quarantine quarantine quarantinelife, nyc lowkey nyc	quarantine stayhome stayhome staysafe, quarantine quarantine stayhome bored, pandemic newyorkcity nyc quarantine, socialdistance nyc stayhome doyourpart, quarantine quarantinelife quarantinelife quarantine,quarantine quarantinelife stayhome staysafe, love socialdistance pandemic nyc, quarantine socialdistance stayhome feetapart, quarantine socialdistance stayhome stircrazy, quarantine stayhome socialdistance wip	quarantine quarantine quarantine quarantine quarantine quarantine quarantine quarantine boy game throne
	Topic 2	china, americans, america, chinese, obama, biden, american, hoax, election, blame, lie, dead, vote, believe	trump china, china trump, trump obama, trump america, trump americans, america trump, trump chinese, obama trump, hoax trump, biden trump	president trump china, china president trump, donald trump china, trump china response, ask china trump, trump ask china, blame china trump, trump trust china, president trump chinese, china trump lose	lie trump president trump, trump biden trump biden, trump america maga trump, trump america save trump, trump supporter believe trump, trump praise china china, lab china president trump, china option president trump, president trump beat china, president trump blame china	presidential race head head biden trump battleground trump biden handle economy trump biden handle health care biden trump handling
	Topic 3	white_house, anthony_fauci, task_force, rand_paul, penny, greta_thunberg, fauci, cdc, donald, administration, cnn, staff, senate, official, negative, aide, result, warn, pence	trump fauci, trump pence, pence trump, donald trump, trump administration, negative trump, trump penny, penny trump, fauci trump, administration trump	trump white house, white house trump, president donald trump, president trump pence, trump pence test, trump fauci penny, penny trump fauci, trump fauci warn, trump fauci play, trump fauci warning	white house president trump, idiot trump trump penny, trump trump fauci present, trump pence test positive, trump administration trump gop, trump administration trump create, white house trump remark, trump pence white house, trump aide infect trump, white house doctor trump	vice president mike pence press secretary test positive donald trump personal valet test positive
	Topic 4	season, travel, sport, college, summer, game, cancel, event, play, flight, baseball, player, fan, class, football, mlb, airline, postpone, concert, graduation, fall, refund, wedding, fly, ruin	season sport, sport season, season team, nba season, school sport, travel school, season player	spring sport season, high school sport, trump nfl season, pandemic college sport, pandemic season game, season travel europe, fall sport season, nba season cancel, future travel travel, pandemic sport event	high school sport season, njsiaa spring sport season, cancel spring season sport, cancel spring sport season, spring high school sport, high school spring sport, high school baseball season, high school college student, future high school sport, sport recent major sport	nba season suspendednhl season suspendedmlb season delay daysus government self quarantine affect
Period-2	Topic 1	public, business, support, vaccine, risk, community, high, patient, crisis, study, impact, learn, flu, result	public health, health public, business health, risk health, health risk, community health, health community, health study, health crisis, patient health	pandemic public health, nyc public health, public health big, vaccine public health, support mental health, public health business, public health infrastructure, public health community, work public health, public health emergency	health care public health, impact health health care, community health mobile health, global public health pandemic, pandemic public health crisis, public health pandemic crisis, pandemic public health expert, public health care crisis, support health care worker, pandemic community health center	pay exam pay essay pay stat pay assignment pay paper pay final paper research pay homework pay economic pay finance pay accounting pay economic pay project pay online class pay business
	Topic 2	chicken, food, eat, pizza, dinner, bread, cake, lunch, cooking, meat, taste, cook, fat, ready, finger_lickin, karen, ugh, kfc, dip, thread, agree, boo, pound, yikes, sigh, slogan, nan	love chicken, fried chicken, eat chicken, love food, chicken sandwich, pizza chicken, chicken salad, chicken dinner, homemade chicken	cream cheese chicken, chicken chicken riggie, love cooking chicken, fast food chicken, love chicken beef, chicken cheese steak, eat fry chicken, love blue chicken, butter chicken pizza, fried chicken pizza	chicken grill chicken rice, fry chicken chicken thigh, love fried chicken universal, chicken cheese steak fry, menu chicken cheese steak, chicken bread chicken cutlet, chicken tender fast food, dinner chicken alfredo dinner, chicken cheese steak sell	addict gummy bears tacos raw cookie dough burritos nachos cake pie popsicle fry chicken cheesecake sour patch kid reese peanut butter cup string cheese sour gummy worm oreo cookies pizza fruit snacks cheese wheels rotisserie chickens
	Topic 3	trump, biden, president, americans, america, china, vote, dead, election, country, economy, lie, rnc, protest, kill, believe, hoax, blame	president trump, biden trump, trump biden, trump americans, americans trump, trump china, trump president, trump desperate, trump america	trump president trump, president trump trump, trump america trump, trump china trump, donald trump trump, trump hoax trump, hurt trump trump, trump election trump, trump country trump	blame trump trump trump, trump trump trump insert, biden trump biden trump, president trump president trump, donald trump president trump, donald trump america trump, face president trump trump, biden trump hillary trump, hydroxychloroquine trump trump biden, delusional trump president trump	donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail donald trump vote mail
Period-3	Topic 1	americans, america, dead, china, american, lie, kill, save, blame	americans trump, trump americans, american trump, trump dead, trump america, dead trump, trump china, trump american, lie trump, america trump	trump trump lie, blame trump trump, trump americans dead, trump america biden, americans dead trump, trump america americans, dead americans trump, biden trump kill, americans infect trump, americans trump responsible	america trump trump incourage, trump lie trump promise, americans dead president trump, donald trump president americans, dead americans child trump, biden blame president trump, trump biden usa america, americans dead elect trump, matter biden trump america, donald trump dead americans	trump hate group stand trump unfit president trump violent action hate group votetrumpout large number trumpisaracist americans dead
	Topic 2	social_distancing, super_spreader, social_distance, wear, rally, event, distance, wearamask, crowd, party, sick, stupid, protect, celebration, idiot, street, inauguration, joke, forget	trump wear, rally trump, trump rally, wear mask, event trump, mask wear, trump event, rally mask, mask rally, sick trump	wear mask trump, trump wear mask, trump rally trump, wear mask mask, trump rally mask, mask wear mask, wear mask wear, social distancing trump, wear mask rally, biden wear mask	trump rally trump rally, trump rally trump supporter, wear mask wear mask, wear mask trump rally, trump lie wear mask, melania trump wear mask, afraid trump wear mask, protect trump wear mask, recklessly mask trump rally, trump supporter wear mask	trump refuse wear mask trump hold indoor rally
	Topic 3	pfizer_biontech, chief_staff, mark_meadow, pfizer, ben_carson, percent_effective, effective, hud_secretary, carson, operation_warp, result, break, trial, early, speed, negative, analysis, candidate, develop, disappear	trump pfizer, pfizer vaccine, vaccine pfizer, pfizer trump, pfizer positive, pfizer early, news pfizer, vaccine trial, early vaccine, president pfizer	pfizer vaccine pfizer, vaccine pfizer early, vaccine news pfizer, vaccine effective trump, vaccine effective pfizer, pfizer data vaccine, pfizer news trump, pfizer vaccine news, positive vaccine result, vaccine pfizer drug	pfizer vaccine effective trump, vaccine trial pfizer vaccine, trump chief staff positive, vaccine pfizer early signal, break trump positive test, vaccine candidate effective pfizer, break news president trump, vaccine ready pfizer effective, vaccine pfizer early peek, trump chief staff meadow	trump administration disaster white house chief staff mark meadows trump aide test positive

Lebanon					
Topic	Keywords	Phrases-L2	Phrases-L3	Phrases-L4	Phrases-Lmax
Topic 1	عراق, حدود, مقتدى, اغلقوا, عراقي, النجف, لعراق, عراقين	فايروس عراق, عراق ايران, عراق ساحه, موصل عراق, عراق شاه, لايدخل عراق, جنوب عراق, عراق همتكم, غلقنا حدود, لعراق ضحك	الصين يهد عراق, ينتفض عراق ينتفض, طيران تحالف عراق, عراق عايش قدره, عراق جيب هالويه, غلقو مطارات حدود, وصل لعراق ثره, يريد يجون لعراق, عراقي ومستعد انكلم, حديث عراقين فهذي	وباء عراق شعب عراق, قادم عراق ينتفض عراق, ارتفاع مصابين فيروس عراق, تعرف فيروس عراق وهم, يسمعلنا اغلقوا حدود ايران, حاله وحيده النجف عراق, عالم امنا عراق يواجه, اصابتهم بفايروس احترازيا عراق, ينتشر عراق اصور الشغب, عراق دكتور اخي متين	طلع مقتدى مقتدى خاتن ثوره مقتدى خاتن لوطن مقتدى الخذر مقتدى مقتدى يصون ثوره هاني رمزي مقتدى يمثل ثوره مقتدى عراق مقتدى يقمع ساحه تحرير
Topic 2	ايغور, مسلم, مليون, مسلمي, ربك, يمهل, عزلت, يمهل	يمهل الصين, مسلم يارب, ربك نسيا, يمهل عقاب	ربك سيا فيروس, يعلم جنود ربك, وكذلك اخذ ربك	ربك نسيا فيروس الصين, الصين تعزل مليون صيني, ربك بظلام لعبيد الصين, قتلت وحبست مليون مسلم ربك نسيا فيروس ,coronarvirusرجل عالم ربك لمرصاد, عظيم ومايعلم جنود ربك, ليعزلوا انفسهم بانفسهم يمهل	الصين الصين الصين تقتل مسلمين الصين تقتل مسلمي ايغور ايغور قضيه مسلمين ايغور تموت براءه وتحيا كلاب الهم
Topic 3	غزه, اسرى, اسرائيلي, اسرائيل, صهيونيه, فلسطيني, شاركونا	قطاع غزه, اصابه غزه, صهيونيه اخطر, واشنطن اسرائيل, سجلت اسرائيل, اسرائيل موقع	يهدك اسرائيل فيروس, رعايه سريريه غزه, يجماعه الخير غزه	عاجل اسرائيل تسجل وفاه, isbackصهيونيه وباء العرب, وزير النقل فلسطيني ندعو	دول عالم تعيش ظروفا حياتيه صعبه بفعل فايروس الم سنّج حجر صحي اجباري حظر لتجو ايقاف لجميع وسائل التنق بين الدول بينما عاني اهالي قطاع غزه الم عاما بفعل حصار فرضه احتلال صهيوني
Topic 4	عصر, اجمالي, متعافين	ينهي عصر, يرفع اجمالي	٠٨٢ اجمالي وفيات, ٢٤٩١ وصل متعافين, ٢١٤٠ وصل متعافين	عالم ٦٨٤ اجمالي مصابين	مارس دول خليج عربي سجلت ١٨٦ اصابه ليصل اجمالي ٢٦٩ اصابه سعوديه ١٠١٢ قطر ٥٤٩ بحرين ٤٥٨ امارات كويت ٢٠٨ سلطنه عمان ١٠٩ حظر تجوال الصوص
Topic 5	تجمعات, بيتك, تجوال, سرعه, غباء, مدن, الحظر	فوائد الحظر	الحظر شامل فيروس, اخوان الحظر جديد, مسؤوليه تزام بيتك	بين مناطق الحظر الكلي, غباء اطلعواوخلوا معزولين لحالهم, الحظر تخاف تطلع تبع	الان متحدث الصحه تجمعات كافه مستوياتها خطره وحتى تجمعات داخل منزل خطره فيروس
Topic 6	تجنب, يدين, غسل, اشخاص, سعال, لمس, اسطح, انف	تعقيم يدين, سعال عطاس, صاب اشخاص, ملامسه اسطح	سعال العطس ذهاب, لمس كمامه خارج	اشخاص بتجيهم اعراض رغم	اهتمامكم وقايه يساعد حمايتكم وحمايه تحبون تجنب مصالحه تجنب اماكن مزدحمه اعزل نفسك مباشره عند وجود اعراض انفلونزا واتصل اهتم بنظافه يدين لانها اكبر ناقل لعدوى مسايقه عنايه وقايه
Topic 7	اسعار, حرب, تجاريه, دولارات	الحرب تجاريه, طلب النفط	-	-	ازمه اغراق السوق نفط مادخل النفط بجانحه هذي جانحه ميه بسبه توقفت مصادر الدخل وتوقفت حركه تجاريه ميا هيا جانحه
Topic 8	موقع, ارمكس, موافع, اكسبرس, ملابس, الشحن, موقوف, شحنه	طلبت طلبيه	-	-	غريبه انك جبت لنا كلامك طريق رابط فلاني ادخلو اكسبرس وحملو منتج
Saudi Arabia					
Topic	**Keywords**	**Phrases-L2**	**Phrases-L3**	**Phrases-L4**	**Phrases-Lmax**
Topic 1	حظرتجول, تعلمت_من_قعدت_البيت, لن_اصافح, خلك_بالبيت	خليك_بالبيت لن_اصافح, حظر_تجول اقعد_في_بيتك, كلمة_الملك_سلمان تعلمت_من_قعدت_البيت, صدق_الله_العظيم	-	-	خارج الصندوق تماما المهم خليك بالبيت لن_اصافح
Topic 2	الصحة, إصابة, متحدث, تعلن, صحيفة_مكة	الصحة تلزم	متحدث الصحة فترة	عاجل متحدث الصحة فايروس, وزارة_الصحة تعلن حالة جديدة, السعودية وزارة الصحة تلزم, اعلنت وزارة الصحة السعودية	عاجل وزارة الصحة السعودية تسجيل إصابة جديدة
Topic 3	الخاص, العمل, القطاع, الموظفين, الشركات, موظفين, قطاع, الحكومي, موظف	القطاع الخاص, بالقطاع الخاص, وزير العمل, الشركات الان, موظفين الصحة	القطاع الخاص ضد, موظفين القطاع الخاص, موظفات القطاع الخاص, تظن القطاع الخاص	موظفين القطاع الخاص يلتزمون, القطاع الخاص الصحي مظلوم, يفرقون موظف القطاع الخاص, الحوامل فقط للقطاع الخاص, القطاع الخاص يتجاهلون القرارات, اغلاق القطاع الخاص مطلب, إلزام القطاعالخاص بتعليق العمل, يفكنا اصحو وزارة العمل, وزارة العمل فضلوها ياخوين	تطلبون منا عزل والبقاء البيوت وانتم ماوقفتو القطاع الخاص العمل ظل الأزمه القطاع الخاص
Topic 4	الطلاب, المسلس, ودون, الطالبات, الطالب, وتتجيهم, تدريب_الامام_كليه_اللغات, واسثناء	وضع الطالب	الأن واسثناء الطلاب, الحل نجحوا الطلاب, تجهيز الطالب نفسيا	تدريب_الامام_كليه_اللغات للأسف مصرين علينا	حضور الطلاب منصة التعليم يفوق الحضور الجسدي الظروف الاعتيادية أحيانا وأيضا تفاعلهم المادة العلمية نعطيها ككادر تعليمي مؤشر لمدى الوعي الجيل الجديد والجميل ومبشر بغد أفضل وأجمل لهذا الوطن
Topic 5	صلوا_في_رحالكم, الجمعه, يوم_الجمعة, الصلاة_في_رحالكم, جمعة_مباركة, تعليق_الصلاه, مختصر_عفيف, الجمعه_الحزينه, جمعة_طيبة, صلاة_الجماعة	صلوا_في_رحالكم صباح_الخير, ساعه_استجابه الجمعه, يوم_الجمعة الجمعه_الحزينه, نصلي الجمعه, يوم_الجمعة صلاة_الجمعة, مختصر_عفيف دعاء, جمعة_طيبة دعاء, الصلاة_في_رحالكم الصلاه_على_النبي, الجمعه_الحزينه المدينة_المنورة	صلوا_في_بيوتكم صلوا_في_رحالكم سنة, هيه_كبار_العلماء الصلاة_في_بيوتكم الصلاة_في_رحالكم, يازين الجمعه الأهل, كلام سليم تعليق_الصلاه	لتقترب جمعة_مباركة جمعة_بالصلاة صباح_الخير	صلاة الوتر آذان أقامة ونقيمها غرفا مادمنا نرجو رحمته ونخشاه وتركم صلوافيرحالكم
Topic 6	فرج, فاتوره, فزعتكم, المبلغ, فرجت_مع_الراوي, دخيل, السداد, كربة, المتبقي	فرج همومنا, كربة كرب, فرجت_مع_الراوي, دخيل الحرم_المكي, فرجت_معا_راوي, دخيل فايز_المالكي, الفاتوره المتبقي	ياأهل الخير فزعتكم, يأهل الخير فزعتكم, الخير ساعدوني لوجه	تقدرون فاتوره فاتوره, يفرج ساهم فرج الام, السداد المبلغ المتبقي ريال, فاتورة التنفيذ المبلغ المتبقي, كربة وأنت رجائي في, نستخرج المبلغ المطلوب بنفس	اللهم فرج أمورا ضاقت صدورنا وعجزت حيلتنا وقل صبرنا اللهم ارزق شارك ساهمو بمانجود انفسكم
Topic 7	شكرا, الوطن, أبطال, أبطال_الصحة, نشكر, أعناق, حقكم, شكرا_أبطال_الصحة, لأبطال, جهودكم	سطر شكرا, وجنود الوطن, السعودية أبطال_الصحة, شكرا_أبطال_الصحة شكرناكم, لأبطال الصحة	شكرا لأبطال الوطن, شكرا أبطال شكرا_أبطال_الصحة, شكرا بحجم السماء, هنيئا أبطال_الصحة الفضل	شكرا أبطال رتويت شكرا_أبطال_الصحة, شكرا ابطال وزارة_الصحة_السعودية وزارة_الصحة, كلمة توفيكم حقكم حافظكم, الف خير جهودكم الجبارة	شكرا وزارة الصحة شكرا أبطال الصحة شكرا أبطال الشرقية
Topic 8	مسلسل, مسلسلات, فلم, أفلام, فيلم, افلام, طاش, المسلسل, مشاهدات	مسلسلات تنفلكس, مسلسل جديد, مسلسل العولمة, مسلسل تركي, مسلسل فاست	فلم افلام تنفلكس, عطوني مسلسل بمستوى, تابعوا مسلسل ماتخلصونه, مسلسل ودي اقضيه	للتو تابعت فيلم فيلم, ومافيه زي مسلسلات زمان, نفسي هالفترة حلقة مسلسل, أتابع مسلسل الإختيار الكويتي, جبت مسلسلات بالتنفلكس مظلومه, تابعتو مسلسل فريندز تابعوه, نصائح ل مشاهدات افلام, فلم الليله جميل فعالبات, تكفون_اقعدوا_في_بيوتكم فلم مطابق لأحداث	اكملت للتو فلم للمخرج مارتن سكورسيزي والمصور روبيرت ريتشاردسون اميسيت ثملا بسحر الفلم منتهى الغموض النهاية متوقعة ابدا أيهما الأفضل
Topic 9	وضع, جنة, الان, الامهات, زلمة, يوم_بس, المتزوجين, الزوجات	الوضع الان, وضع الامهات, وضع البيوت, وضع الميكانيكية, وضع المتقاعدين, وضع العوائل, وضع الأمهات, حال المتزوجين, يستر الجاي	وضع الأباء الان, حال الامهات اطفالهم	افرقنا زلمة معاناة الزوجات, بعدنا بثالث يوم_بس يترى	الزوجين استغلاله وغالبا الأنثى الزوجة تجعل المكان جنة جحيم

Figure 7.4: **Top keywords and phrases extracted using RAKE based on BERTopic top keywords. The keywords and phrases are ranked based on RAKE - Lebanon and Saudi Arabia.**

Chapter 8

Conclusion

8.1 Conclusion and Closing Remarks

This thesis proposes a novel multi-lingual-dialectal framework for modeling multimedia domain-independent online social behavior on social media. We address the limitation of domain dependency in modeling online social behavior and we propose to build domain-independent data resources and models to analyze online social behaviors. Our results confirm the efficacy of the domain-independent approach when our domain-free sentiment model has been able to adapt to different domains (i.e. sports and movie reviews) while domain-dependent models have failed to generalize to general domain. Similar observation has been detected when modeling hate behavior on social media. Without doubt, the informal and slang nature of conversations has become the dominant mode for communication on social media. Not only this, but variety of foreign languages -other than English- have been increasingly used in social talks and conversations. Accordingly, monitoring systems for online social behavior have to extend their ability to understand and analyze interactions expressed in different languages and even dialects. Current resources (i.e. data and models) are not efficiently exploited as researchers tend to provide solutions to a specific problem in a specific language and dialect instead of utilizing the already existing resources especially those in high-resourced languages. Building a new resource itself is challenging and costly in terms of time, efforts, cost, and human labor. It is difficult to have domain experts or workers to commit to properly annotate a large volume of data. The data construction step is very crucial since the learning performance is directly affected if it is done improperly, and this is an ongoing issue that still puzzles researchers up to this moment. As researchers, we are obligated to address the limitations of this current practice and provide reliable and inexpensive solutions to minimize the dependency of languages and/or dialects in modeling online social behaviors on social media. In response, we propose a framework that targets a low-resource language (i.e. Arabic in this thesis) and exploits visual contents as a universal language that could provide insights about online social behaviors. We propose a multi-lingual translation dataset (MLMD) to/from multi-dialectal Arabic in an attempt to contribute in the enrichment of the low-resource status of multi-dialectal Arabic resources. This dataset is then utilized for building neural

machine translation (NMT) models from/to multi languages to/from Arabic multi dialects (Levantine, Gulf, Iraqi, and Yemeni). Note that the proposed NMT models are optimized for OSN data. Our experimental results have illustrated a superior performance of our proposed multi-lingual-dialectal NMT models when compared against Google Translate and two LLM models: GPT3.5 and GPT4. Further, our NMT models have been proven effective as a tool to build OSB models for low-resource languages/dialects as evidenced by our proof-of-concept COVID-19 case study on two low-resourced Arabic dialects (Arabic-Levantine and Arabic-Gulf) where we exploited existing resources in English, French, and Spanish (i.e. high resourced languages). The results have shown that NMT-based OSB models (sentiment and hate in this thesis) yield superior performance in learning all classes for both classification tasks (i.e. sentiment and hate analysis tasks) in both Arabic dialects (Arabic-Levantine and Arabic-Gulf). Not only this, but our OSB models have proven their capability in recognizing local contexts across different dialects of the same language (Arabic in this thesis). This has been shown in Arabic-Levantine hate classifier being able to detect hate contents in data from Lebanon whereas the Arabic-Gulf hate classifier failed to recognize the presence of hate speech from the very same set of data. This finding concludes that OSB models trained for a specific dialect of a language does not perform well on another dialect of the very same language. This actually highlights the importance of the contextualization of communication content for a specific language and its dialect(s). Ignoring this aspect leads to misleading analysis on the underlying online social behavior of OSN communications.

Inspired by emojis that speak a universal language through iconic expressions, we adopt a supplementary language-independent approach where a universal language of visual images and emotions are used to model online social behavior. However, images on social media are uncontrolled, noisy, and freely shared; they are not defined by structured characteristics which, in turn, brings serious challenges in finding a relationship between such images and their emotional/sentimental semantics that are indirectly driven by cognitive semantics. Accordingly, we propose a visual multi-modality classifier that leverages objects in images. Three types of images are considered in this thesis: image with text, image with face, and image with neither face nor text. The corresponding experiments show that exploiting facial emotion (i.e. face object) and textual information (i.e. text object) extracted from images contributes in enhancing the learning of visual online social behavior. Note that sentiment behavior has been considered as a case study to examine our proposed approach.

In order to facilitate the understanding of predicted online social behaviors, we propose using topic modeling and phrase extraction methods for discovering hidden patterns and inferring topics, trends, and concerns as well as automatically providing coherent interpretation of the inferred topics without human effort involved. It is known that unsupervised traditional topic models are sensitive to noise, a natural phenomenon in OSNs data. Accordingly, Natural Language Processing (NLP) techniques have been exploited as prerequirements to topic modeling in order to maximize their learning performance on OSNs data (i.e. English data in this case). Traditional topic models, like LDA and NMF, work well under the condition that data contains the minimum amount of noises. Minimizing or removing noise from data is possible with the availability of reliable preprocessing tools

(e.g. POS, lemmatization, stemming, etc.) that are adequately found in high-resourced languages like English. Such resources are insufficient or even missing in low-resourced languages like Arabic and its dialects, hence, traditional topic models would fail to converge and provide meaningful topics from a given OSN data. While traditional topic learning comes with the advantage that it requires a cheaper hardware resource and performs at a lower computational complexity than deep learning does, deep learning, especially the transformers architecture, omits the pre-requirement of data cleaning/preparation and feature engineering; instead, transformer models learn contextualized features directly from the data while being robust against the noise presence. Since language preprocessing tools to clean data noises are not available for Arabic dialects, deep learning approach has proven to be extremely helpful in exploring multi-dialectal Arabic data even though it requires advanced and expensive hardware resources and performs at a high computational complexity. The results show that both traditional and deep learning topic models produce coherent topics on real data collected in real-time from Twitter. Topic models describe inferred topics through a set of keywords; however, single keywords alone are not enough to help comprehend the main meaning of topics. Sentences also are too specific and might miss other aspects of topics. Conversely, short meaningful phrases provide a clearer interpretation of the topics. Given the diversity of conversations on OSNs, finding the optimal length of phrases that best describe a topic is challenging. Some topics might use longer or shorter phrasal expressions than the others, and this cannot be controlled on open platforms like OSNs, which encourage unstructured data format. This thesis addresses this issue and utilizes the language-independent RAKE algorithm that handles the dynamic-length generation of comprehensible phrases that best describe the inferred topics from OSNs data in different languages and dialects.

A large-scale analysis on COVID-19 data in North America and Middle East has been conducted in this thesis as a proof of concept for our framework. Our exploratory analysis for two far-apart continents confirms the influence of geo-region, culture, and religion on the social behavior of users on the virtual world of social media. It has been observed that religion has a significant impact on culture in the Middle East region. Religion teaches coping mechanisms; this seems to have relaxed the Middle Eastern region during COVID-19 pandemic as the lowest percentage of toxic speech has been found in Saudi Arabia where prayers to Allah has taken a large part of OSN communications. Not only this, but also the sense of community and belonging has been very noticeable through the various campaign initiatives for donation and community support like "Forijat", which is a Saudi service that allows the public to donate funds to help people who have been imprisoned because of their inability to clear debts. On the other hands, political instability seems to have a negative effect on people who inhabit a region. Lebanon has been found to not only have a high amount of sad vibes but also to score the highest percentage of toxic behavior during COVID-19 pandemic. This is possibly not due to the pandemic alone but it is also a protest against the wretched economic situation over there. Further, North America seems to have been conservative and concerned over the local issues related to the north continent like elections and public health. "Trump" related talks, for instance, seems to have been opinionated in North America, while it has been mostly news reports in the Middle East region.

8.2 Future Directions

While this thesis demonstrates the feasibility of multi-lingual-dialectal online social behavior modeling, it can be extended and improved in several ways:

- Targeting dialects of other languages: As discussed in Chapter 5, dialects within the same language are different; adopting a dialectal model on another different dialect leads to false OSB detection. Languages like Spanish has many dialects including Argentinian, Spain-Spanish, Ecuadorian, and Mexican. Therefore, an immediate extension to this thesis would be to develop OSB models that capture the local sense of these dialects in OSN communications. Unfortunately, dialectal resources for many languages are insufficient or almost none-existent.

- Consideration of videos and speech for multi-lingual-dialectal OSB modeling: Video and voice OSN communications contain an immense amount of dialectal contents that have not been well exploited for OSB analysis on social media. Integrating visual, speech, and textual information to model multi-lingual-dialectal OSB would lead to enriched analysis of social behaviors that are being increasingly generated over social media.

- Consideration of ontology approach for OSB modeling: OSN data is different from any other types of data; it does not follow a unified structure as OSN platforms encourage free sharing of contents. It is challenging to find a relationship between such data and emotional/sentimental semantics (i.e. especially in visual content) as they are driven indirectly by cognitive semantics. It is near to impossible to control the content generated on social media; however, it is possible to control the structure of the cognitive component and customize it to OSB modeling. This would alleviate the computational complexity of deep learning methods and the main dependency of feature engineering, and hence domain experts in traditional machine learning approach.

- Modeling and assessment of other online social behaviors: This thesis studies sentiment and toxic speech behaviors. We believe that expanding the set of online social behaviors to include emotion, sarcasm, and optimism (to name a few) would provide access to multiple views and perspectives of citizens' behaviors within and across smart cities.

- Construction of larger data resources: collecting and expanding our existing data resources are necessary to advance this work and expand its applications. We plan to enrich the under-resource status of Arabic dialects by expanding our MT dataset to include the rest of the dialects that we have not targeted in this thesis. We also plan to expand our Spanish and French MT datasets to include the different dialects of Spanish and French.

- Generation of automatic multimedia stories: This thesis provides temporal, spatial, cultural, and topic-based views of OSB analysis on social media. However, creating

multimedia stories at the current stage of this work requires handling the information overload of textual and visual content [16] in order to choose a set of the most representative messages, images, topics that best describe each view. In the future, we plan to use certain techniques to estimate how many messages and images per view will be used and which messages and images will represent each view in order to automatically generate enriched multimedia stories that reflect the sense of smart cities from different angles.

References

[1] Rana Abaalkhail, Fatimah Alzamzami, Samah Aloufi, Rajwa Alharthi, and Abdulmotaleb El Saddik. Affectional ontology and multimedia dataset for sentiment analysis. In *International Conference on Smart Multimedia*, pages 15–28. Springer, 2018.

[2] Muhammad Abdul-Mageed and Mona T Diab. Sana: A large scale multi-genre, multi-dialect lexicon for arabic subjectivity and sentiment analysis. In *LREC*, pages 1162–1169, 2014.

[3] Abeer Abuzayed and Hend Al-Khalifa. Bert for arabic topic modeling: An experimental study on bertopic technique. *Procedia computer science*, 189:191–194, 2021.

[4] Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. Detecting white supremacist hate speech using domain specific word embedding with deep learning and bert. *IEEE Access*, 9:106363–106374, 2021.

[5] Sahar Aldhaheri. Sentiment analysis for saudi public opinion toward covid19 and quarantine. https://www.linkedin.com/pulse/sentiment-analysis-saudi-public-opinion-toward-sahar-aldhaheri/, 2020. Accessed: (August 23, 2023).

[6] Rajwa Alharthi, Benjamin Guthier, Camille Guertin, and Abdulmotaleb El Saddik. A dataset for psychological human needs detection from social networks. *IEEE Access*, 5:9109–9117, 2017.

[7] Samah Aloufi, Fatimah Alzamzami, Mohamad Hoda, and Abdulmotaleb El Saddik. Soccer fans sentiment through the eye of big data: The uefa champions league as a case study. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 244–250. IEEE, 2018.

[8] Samah Aloufi and Abdulmotaleb El Saddik. Sentiment identification in football-specific tweets. *IEEE Access*, 6:78609–78621, 2018.

[9] Dhuha Alqahtani, Lama Alzahrani, Maram Bahareth, Nora Alshameri, Hend Al-Khalifa, and Luluh Aldhubayi. Customer sentiments toward saudi banks during the covid-19 pandemic. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 251–257, 2022.

[10] Meshrif Alruily and Osama R Shahin. Sentiment analysis of twitter data for saudi universities. *International Journal of Machine Learning and Computing*, 10(1), 2020.

[11] Ali Alshahrani, Meysam Ghaffari, Kobra Amirizirtol, and Xiuwen Liu. Identifying optimism and pessimism in twitter messages using xlnet and deep consensus. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[12] Ali Alshahrani, Meysam Ghaffari, Kobra Amirizirtol, and Xiuwen Liu. Optimism/pessimism prediction of twitter messages and users using bert with soft label assignment. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[13] Raghad Alshalan and Hend Al-Khalifa. A deep learning approach for automatic hate speech detection in the saudi twittersphere. *Applied Sciences*, 10(23):8614, 2020.

[14] Raghad Alshalan, Hend Al-Khalifa, Duaa Alsaeed, Heyam Al-Baity, and Shahad Alshalan. Hate detection in covid-19 tweets in the arab region using deep learning and topic modeling. *Journal of Medical Internet Research*, 2020.

[15] Sarah N Alyami and Sunday O Olatunji. Application of support vector machine for arabic sentiment classification using twitter-based dataset. *Journal of Information & Knowledge Management*, 19(01):2040018, 2020.

[16] Fatimah Alzamzami. *Towards Multimedia-Based Storytelling in Online Social Networks*. PhD thesis, Université d'Ottawa/University of Ottawa, 2015.

[17] Fatimah Alzamzami and Abdulmotaleb El Saddik. Monitoring cyber sentihate social behavior during covid-19 pandemic in north america. *IEEE Access*, 2021.

[18] Fatimah Alzamzami and Abdulmotaleb El Saddik. Transformer-based feature fusion approach for multimodal visual sentiment recognition using tweets in the wild. *IEEE Access*, 2023.

[19] Fatimah Alzamzami, Mohamad Hoda, and Abdulmotaleb El Saddik. Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation. *IEEE Access*, 2020.

[20] Fatimah Alzamzami and Abdulmotaleb El Saddik. Content-localization based neural machine translation for informal dialectal arabic: Spanish/french to levantine/gulf arabic, 2023.

[21] Fatimah Alzamzami and Abdulmotaleb El Saddik. Content-localization based system for analyzing sentiment and hate behaviors in low-resource dialectal arabic: English to levantine and gulf. *arXiv preprint arXiv:2312.03727*, 2023.

[22] Fatimah Alzamzami, Mukesh Saini, and Abdulmotaleb El Saddik. Dst: days spent together using soft sensory information on osns—a case study on facebook. *Soft Computing*, 21(15):4227–4238, 2017.

[23] Mayank Amencherla and Lav R Varshney. Color-based visual sentiment for social communication. In *2017 15th Canadian Workshop on Information Theory (CWIT)*, pages 1–5. IEEE, 2017.

[24] Syed Umar Amin et al. Deep learning for eeg motor imagery classification based on multi-layer cnns feature fusion. *Future Generation Computer Systems*, 101:542–554, 2019.

[25] Muhammad Zubair Asghar, Fazal Masud Kundi, Shakeel Ahmad, Aurangzeb Khan, and Furqan Khan. T-saf: Twitter sentiment analysis framework using a hybrid classification scheme. *Expert Systems*, 35(1):e12233, 2018.

[26] Vasileios Athanasiou and Manolis Maragoudakis. A novel, gradient boosting framework for sentiment analysis in languages where nlp resources are not plentiful: a case study for modern greek. *Algorithms*, 10(1):34, 2017.

[27] Noureddine Azzouza, Karima Akli-Astouati, and Roliana Ibrahim. Twitterbert: Framework for twitter sentiment analysis based on pre-trained language model representations. In *International Conference of Reliable Information and Communication Technology*, pages 428–437. Springer, 2019.

[28] Alexandra Balahur and Marco Turchi. Multilingual sentiment analysis using machine translation? In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 52–60, 2012.

[29] Ramy Baly, Georges El-Khoury, Rawan Moukalled, Rita Aoun, Hazem Hajj, Khaled Bashir Shaban, and Wassim El-Hajj. Comparative evaluation of sentiment analysis methods across arabic dialects. *Procedia Computer Science*, 117:266–273, 2017.

[30] Ramy Baly, Alaa Khaddaj, Hazem Hajj, Wassim El-Hajj, and Khaled Bashir Shaban. Arsentd-lev: A multi-topic corpus for target-based sentiment analysis in arabic levantine tweets. *arXiv preprint arXiv:1906.01830*, 2019.

[31] Laith H Baniata, Seyoung Park, and Seong-Bae Park. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational intelligence and neuroscience*, 2018, 2018.

[32] Amparo Elizabeth Cano Basave, Yulan He, and Ruifeng Xu. Automatic labelling of topic models learned from twitter by summarisation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 618–624, 2014.

[33] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics, 2019.

[34] Fabian Beijer. The syntax and pragmatics of exclamations and other expressive/emotional utterances. *Working Papers in Linguistics*, 2, 2002.

[35] Muhammad Bilal, Huma Israr, Muhammad Shahid, and Amin Khan. Sentiment classification of roman-urdu opinions using naïve bayesian, decision tree and knn classification techniques. *Journal of King Saud University-Computer and Information Sciences*, 28(3):330–344, 2016.

[36] David M Blei and John D Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10(71):34, 2009.

[37] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[38] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[39] Monali Bordoloi and Saroj Kumar Biswas. Sentiment analysis: A survey on design framework, applications and future scopes. *Artificial Intelligence Review*, pages 1–56, 2023.

[40] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232, 2013.

[41] Houda Bouamor, Nizar Habash, and Kemal Oflazer. A multidialectal parallel corpus of arabic. In *LREC*, pages 1240–1245, 2014.

[42] Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[43] Mondher Bouazizi and Tomoaki Ohtsuki. A pattern-based approach for multi-class sentiment analysis in twitter. *IEEE Access*, 5:20617–20639, 2017.

[44] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[45] Ioan Buciu and Ioannis Pitas. Application of non-negative and local non negative matrix factorization to facial expression recognition. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 1, pages 288–291. IEEE, 2004.

[46] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[47] Sean T Campbell and Dattesh R Dave. Two sides to every story: early specialization in medical education. *Medical Science Educator*, 28:243–246, 2018.

[48] Victor Campos, Brendan Jou, and Xavier Giro-i Nieto. From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. *Image and Vision Computing*, 65:15–22, 2017.

[49] Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461, 2017.

[50] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[51] Yuxin Chen, Jean-Baptiste Bordes, and David Filliat. An experimental comparison between nmf and lda for active cross-situational object-word learning. In *2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 217–222. IEEE, 2016.

[52] Raphael Cohen-Almagor. Fighting hate and bigotry on the internet. *Policy & Internet*, 3(3):1–26, 2011.

[53] Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22, 2020.

[54] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005.

[55] Soheil Danesh, Tamara Sumner, and James H Martin. Sgrank: Combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction. In *Proceedings of the fourth joint conference on lexical and computational semantics*, pages 117–126, 2015.

[56] Yan Dang, Yulei Zhang, and Hsinchun Chen. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4):46–53, 2010.

[57] William M Darling, Michael Paul, and Fei Song. Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. In *Proceedings of the Workshop on Semantic Analysis in Social Media*, pages 1–9, 2012.

[58] Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4):757–771, 2016.

[59] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Coling 2010: Posters*, pages 241–249, 2010.

[60] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[61] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[62] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[63] Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. Adversarial and domain-aware bert for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, 2020.

[64] Christopher Dyer, Aaron Cordova, Alex Mont, and Jimmy Lin. Fast, easy, and cheap: Construction of statistical machine translation models with mapreduce. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 199–207. Association for Computational Linguistics, 2008.

[65] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587, 2011.

[66] Angela Fan, Finale Doshi-Velez, and Luke Miratrix. Promoting domain-specific terms in topic models with informative priors. *arXiv preprint arXiv:1701.03227*, 2017.

[67] Angela Fan, Finale Doshi-Velez, and Luke Miratrix. Assessing topic model relevance: Evaluation and informative priors. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):210–222, 2019.

[68] Angela Fan, Finale Doshi-Velez, and Luke Miratrix. Assessing topic model relevance: Evaluation and informative priors. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):210–222, 2019.

[69] Minghui Fan, Andrew Billings, Xiangyu Zhu, and Panfeng Yu. Twitter-based birging: Big data analysis of english national team fans during the 2018 fifa world cup. *Communication & Sport*, page 2167479519834348, 2019.

[70] Dalya Faraj and Malak Abdullah. Sarcasmdet at sarcasm detection task 2021 in arabic using arabert pretrained model. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 345–350, 2021.

[71] Ibrahim Abu Farha and Walid Magdy. From arabic sentiment analysis to sarcasm detection: The arsarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 32–39, 2020.

[72] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[73] Xiaoyi Feng, M Pietikainen, and Abdenour Hadid. Facial expression recognition with local binary patterns and linear programming. *Pattern Recognition And Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii*, 15(2):546, 2005.

[74] JIN Fenglei, Gao Cuiyun, et al. An online topic modeling framework with topics automatically labeled. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 73–76, 2019.

[75] Corina Florescu and Cornelia Caragea. Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1115, 2017.

[76] Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12, 2018.

[77] Keita Fujihira and Noriko Horibe. Multilingual sentiment analysis for web text based on word to word translation. In *2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 74–79, 2020.

[78] Frank Fujita, Ed Diener, and Ed Sandvik. Gender differences in negative affect and well-being: the case for emotional intensity. *Journal of personality and social psychology*, 61(3):427, 1991.

[79] Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. Target-dependent sentiment classification with bert. *IEEE Access*, 7:154290–154299, 2019.

[80] Manoochehr Ghiassi and S Lee. A domain transferable lexicon set for twitter sentiment analysis using a supervised machine learning approach. *Expert Systems with Applications*, 106:197–216, 2018.

[81] Anastasia Giachanou and Fabio Crestani. Like it or not: A survey of twitter sentiment analysis methods. *ACM Computing Surveys (CSUR)*, 49(2):28, 2016.

[82] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

[83] Claudia Goldin. The human-capital century and american leadership: Virtues of the past. *The Journal of Economic History*, 61(2):263–292, 2001.

[84] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pages 117–124. Springer, 2013.

[85] Antoine Gourru, Julien Velcin, Mathieu Roche, Christophe Gravier, and Pascal Poncelet. United we stand: Using multiple strategies for topic labeling. In *International Conference on Applications of Natural Language to Information Systems*, pages 352–363. Springer, 2018.

[86] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022.

[87] Olivier Habimana, Yuhua Li, Ruixuan Li, Xiwu Gu, and Ge Yu. Sentiment analysis using deep learning approaches: an overview. *Science China Information Sciences*, 63(1):1–36, 2020.

[88] Hatem Haddad, Hala Mulki, and Asma Oueslati. T-hsab: A tunisian hate speech and abusive dataset. In *International Conference on Arabic Language Processing*, pages 251–263. Springer, 2019.

[89] Hadry. French twitter sentiment analysis. https://kaggle.com/hbaflast/french-twitter-sentiment-analysis, 2020. Accessed: (August 15, 2023).

[90] Gerhard J Hanneman. The study of human communication. *Communication and behavior*, pages 21–49, 1975.

[91] Kazi Saidul Hasan and Vincent Ng. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273, 2014.

[92] Dongbin He, Minjuan Wang, Abdul Mateen Khattak, Li Zhang, and Wanlin Gao. Automatic labeling of topic models using graph-based ranking. *IEEE Access*, 7:131593–131608, 2019.

[93] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[94] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[95] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

[96] Jayme Hill Hill. The impact of emojis and emoticons on online consumer reviews, perceived company response quality, brand relationship, and purchase intent. 2016.

[97] Kalun Ho. Multilingual sentiment analysis on social media using deep learning, 2017.

[98] M. Shamim Hossain et al. Audio–visual emotion-aware cloud gaming framework. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(12):2105–2118, 2015.

[99] M. Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78, 2019.

[100] Binxuan Huang and Kathleen Carley. Parameterized convolutional neural networks for aspect level sentiment classification. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1091–1096, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[101] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[102] Qionghao Huang, Changqin Huang, Xizhe Wang, and Fan Jiang. Facial expression recognition with grid-wise attention and visual transformer. *Information Sciences*, 580:35–54, 2021.

[103] Google Icn. *Google Translator*, (accessed February 20, 2024).

[104] Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, and Rakhi Batra. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. *IEEE Access*, 8:181074–181090, 2020.

[105] Mohammed Jabreel and Antonio Moreno. EiTAKA at SemEval-2018 task 1: An ensemble of n-channels ConvNet and XGboost regressors for emotion analysis of tweets. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors, *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 193–199, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[106] Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522, 2019.

[107] Byeongki Jeong, Janghyeok Yoon, and Jae-Min Lee. Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis. *International Journal of Information Management*, 48:280–290, 2019.

[108] Mengxiao Jiang, Man Lan, and Yuanbin Wu. Ecnu at semeval-2017 task 5: An ensemble of regression algorithms with effective features for fine-grained sentiment analysis in financial domain. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 888–893, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[109] Stuti Jindal and Sanjay Singh. Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning. In *2015 International Conference on Information Processing (ICIP)*, pages 447–451. IEEE, 2015.

[110] Vineet John and Olga Vechtomova. UW-FinSent at SemEval-2017 task 5: Sentiment analysis on financial news headlines using training dataset augmentation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 872–876, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[111] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, pages 3146–3154, 2017.

[112] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2, 2019.

[113] Bo-Kyeong Kim, Jihyeon Roh, Suh-Yeon Dong, and Soo-Young Lee. Hierarchical committee of deep convolutional neural networks for robust facial expression recognition. *Journal on Multimodal User Interfaces*, 10(2):173–189, 2016.

[114] Olga Kolchyna, Tharsis TP Souza, Philip Treleaven, and Tomaso Aste. Twitter sentiment analysis: Lexicon method, machine learning method and their combination. *arXiv preprint arXiv:1507.00955*, 2015.

[115] Puneet Kumar, Vedanti Khokher, Yukti Gupta, and Balasubramanian Raman. Hybrid fusion based approach for multimodal emotion recognition with insufficient labeled data. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 314–318, 2021.

[116] Mateusz Lango, Dariusz Brzezinski, and Jerzy Stefanowski. Put at semeval-2016 task 4: The abc of twitter sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 126–132, 2016.

[117] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, 2014.

[118] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[119] Diya Li, Harshita Chaudhary, and Zhe Zhang. Modeling spatiotemporal pattern of depressive symptoms caused by covid-19 using social media data mining. *International Journal of Environmental Research and Public Health*, 17(14):4988, 2020.

[120] Hanting Li, Mingzhe Sui, Feng Zhao, Zhengjun Zha, and Feng Wu. Mvit: Mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*, 2021.

[121] Lingxiao Li, Shaozi Li, Donglin Cao, and Dazhen Lin. Sentinet: Mining visual sentiment from scratch. In *Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK*, pages 309–317. Springer, 2017.

[122] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[123] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[124] Zhongli Li, Shiai Zhu, Huiwen Hong, Yuanyuan Li, and Abdulmotaleb El Saddik. City digital pulse: a cloud based heterogeneous data analysis platform. *Multimedia Tools and Applications*, 76(8):10893–10916, 2017.

[125] Zuhe Li, Yangyu Fan, Bin Jiang, Tao Lei, and Weihua Liu. A survey on sentiment analysis and opinion mining for social multimedia. *Multimedia Tools and Applications*, 78:6939–6967, 2019.

[126] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[127] Jimmy Lin and Alek Kolcz. Large-scale machine learning at twitter. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 793–804. ACM, 2012.

[128] Bing Liu et al. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2(2010):627–666, 2010.

[129] Ning Liu, Benyu Zhang, Jun Yan, Zheng Chen, Wenyin Liu, Fengshan Bai, and Leefeng Chien. Text representation: From vector to tensor. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 4–pp. IEEE, 2005.

[130] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[131] Fuyan Ma, Bin Sun, and Shutao Li. Facial expression recognition with visual transformers and attentional selective fusion. *IEEE Transactions on Affective Computing*, 2021.

[132] Fuyan Ma, Bin Sun, and Shutao Li. Robust facial expression recognition with convolutional visual transformers. *arXiv preprint arXiv:2103.16854*, 2021.

[133] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

[134] Kaleel Mahmood, Rigel Mahmood, and Marten Van Dijk. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7838–7847, 2021.

[135] Ofer Malamud. Breadth vs. depth: The timing of specialization in higher education. nber working paper no. 15943. *National Bureau of Economic Research*, 2010.

[136] Xian-Ling Mao, Yi-jing Zhao, Qiang Zhou, Wen-Qing Yuan, Liner Yang, and He-Yan Huang. A novel fast framework for topic labeling based on similarity-preserved hashing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3339–3348, 2016.

[137] Fiona Martin and Mark Johnson. More efficient topic modelling through a noun only approach. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 111–115, 2015.

[138] Qiaozhu Mei, Xuehua Shen, and ChengXiang Zhai. Automatic labeling of multinomial topic models. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 490–499, 2007.

[139] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12309–12318, 2022.

[140] Yu Miao, Haiwei Dong, Jihad Mohamad Al Jaam, and Abdulmotaleb El Saddik. A deep learning system for recognizing facial expression in real-time. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2):1–20, 2019.

[141] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[142] Saif Mohammad and Felipe Bravo-Marquez. Emotion intensities in tweets. In Nancy Ide, Aurélie Herbelot, and Lluís Màrquez, editors, *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 65–77, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[143] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In Suresh Manandhar and Deniz Yuret, editors, *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[144] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[145] Wajid Hassan Moosa and Najiba. Multi-lingual hatespeech dataset, 2022.

[146] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer, 2019.

[147] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. *PloS one*, 15(8):e0237861, 2020.

[148] Aminu Muhammad, Nirmalie Wiratunga, Robert Lothian, and Richard Glassey. Domain-based lexicon enhancement for sentiment analysis. In *SMA@ BCS-SGAI*, pages 7–18, 2013.

[149] Sandeep Sricharan Mukku, Subba Reddy Oota, and Radhika Mamidi. Tag me a label with multi-arm: Active learning for telugu sentiment analysis. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 355–367. Springer, 2017.

[150] Hala Mulki and Bilal Ghanem. Let-mi: An Arabic Levantine Twitter dataset for misogynistic language. In Nizar Habash, Houda Bouamor, Hazem Hajj, Walid Magdy, Wajdi Zaghouani, Fethi Bougares, Nadi Tomeh, Ibrahim Abu Farha, and Samia Touileb, editors, *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 154–163, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics.

[151] Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online*, pages 111–118, 2019.

[152] Manish Munikar, Sushil Shakya, and Aakash Shrestha. Fine-grained sentiment classification using bert. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–5. IEEE, 2019.

[153] Batsergelen Myagmar, Jie Li, and Shigetomo Kimura. Transferable high-level representations of bert for cross-domain sentiment classification. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, pages 135–141. The Steering Committee of The World Congress in Computer Science, Computer . . . , 2019.

[154] Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. SemEval-2013 task 2: Sentiment analysis in Twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[155] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[156] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449, 2015.

[157] Minh-Tien Nguyen, Dac Viet Lai, Huy Tien Nguyen, and Minh Le Nguyen. Tsix: a human-involved-creation dataset for tweet summarization. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[158] Thien Hai Nguyen and Kiyoaki Shirai. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1354–1364, 2015.

[159] Finn Årup Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[160] Alex Nikolov and Victor Radivchev. Nikolov-radivchev at semeval-2019 task 6: Offensive tweet classification with bert and ensembles. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 691–695, 2019.

[161] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. Sentiment analysis on multi-view social data. In *International Conference on Multimedia Modeling*, pages 15–27. Springer, 2016.

[162] Diogo Nolasco and Jonice Oliveira. Subevents detection through topic modeling in social media posts. *Future Generation Computer Systems*, 93:290–303, 2019.

[163] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS one*, 10(12):e0144296, 2015.

[164] Committee of Ministers. *Council of Europe*, 1997 (accessed December 1, 2020).

[165] Marwan Al Omari. Oclar: logistic regression optimisation for arabic customers' reviews. *International Journal of Business Intelligence and Data Mining*, 20(3):251–273, 2022.

[166] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu,

Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

[167] Catherine Ordun, Sanjay Purushotham, and Edward Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. *arXiv preprint arXiv:2005.03082*, 2020.

[168] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato. Survey on visual sentiment analysis. *IET Image Processing*, 14(8):1440–1456, 2020.

[169] Mathieu Pagé Fortin and Brahim Chaib-draa. Multimodal multitask emotion recognition using images, texts and tags. In *Proceedings of the ACM Workshop on Cross-*

modal Learning and Application, pages 3–10, 2019.

[170] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[171] Endang Wahyu Pamungkas and Viviana Patti. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, 2019.

[172] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[173] Eirini Papagiannopoulou and Grigorios Tsoumakas. A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(2):e1339, 2020.

[174] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[175] Alberto Poncelas, Pintu Lohar, Andy Way, and James Hadley. "the impact of indirect machine translation on sentiment classification". In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 78–85, Virtual, October 2020. Association for Machine Translation in the Americas.

[176] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[177] Kumar Ravi and Vadlamani Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, 2015.

[178] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France, May 2020. European Language Resources Association.

[179] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In Nicoletta Calzolari, Frédéric Béchet, Philippe

Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France, May 2020. European Language Resources Association.

[180] Marko Robnik-Šikonja. Improving random forests. In *European conference on machine learning*, pages 359–370. Springer, 2004.

[181] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1:1–20, 2010.

[182] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518, 2017.

[183] Marta Ruiz Costa-Jussà, Marcos Zampieri, and Santanu Pal. A neural approach to language variety translation. In *COLING 2018: The 27th International Conference on Computational Linguistics: Proceedings of the Conference: August 20-26, 2018 Santa Fe, New Mexico, USA*. Association for Computational Linguistics, 2018.

[184] Mukesh Kumar Saini, Fatimah Al-Zamzami, and Abdulmotaleb El Saddik. Towards storytelling by extracting social information from osn photo's metadata. In *Proceedings of the First International Workshop on Internet-Scale Multimedia Management*, pages 15–20, 2014.

[185] Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. Arabench: Benchmarking dialectal arabic-english machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, 2020.

[186] Hassan Sajjad, Nadir Durrani, Francisco Guzman, Preslav Nakov, Ahmed Abdelali, Stephan Vogel, Wael Salloum, Ahmed El Kholy, and Nizar Habash. Egyptian arabic to english statistical machine translation system for nist openmt'2015. *arXiv preprint arXiv:1606.05759*, 2016.

[187] Wael Salloum and Nizar Habash. Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties*, pages 10–21, 2011.

[188] Wael Salloum and Nizar Habash. Dialectal arabic to english machine translation: Pivoting through modern standard arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358, 2013.

[189] Niloofar Safi Samghabadi, Parth Patwa, PYKL Srinivas, Prerana Mukherjee, Amitava Das, and Thamar Solorio. Aggression and misogyny detection using bert: A

multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, 2020.

[190] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[191] Michael B Schiffer. Archaeological context and systemic context. *American antiquity*, 37(2):156–165, 1972.

[192] Michael Brian Schiffer. *The material life of human beings: artifacts, behavior and communication*. Routledge, 2002.

[193] Alexandra Schofield, Måns Magnusson, and David Mimno. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436, 2017.

[194] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Robust facial expression recognition using local binary patterns. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–370. IEEE, 2005.

[195] Pamela Shapiro and Kevin Duh. Comparing pipelined and integrated approaches to dialectal arabic neural machine translation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 214–222, 2019.

[196] Anuj Sharma and Shubhamoy Dey. A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM research in applied computation symposium*, pages 1–7. ACM, 2012.

[197] Zeeshan Shaukat, Abdul Ahad Zulfiqar, Chuangbai Xiao, Muhammad Azeem, and Tariq Mahmood. Sentiment analysis on imdb using lexicon and neural networks. *SN Applied Sciences*, 2:1–10, 2020.

[198] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1105–1114, 2018.

[199] Sifatullah Siddiqi and Aditi Sharan. Keyword and keyphrase extraction techniques: a literature review. *International Journal of Computer Applications*, 109(2), 2015.

[200] Stefan Siersdorfer, Enrico Minack, Fan Deng, and Jonathon Hare. Analyzing and predicting sentiment of images on the social web. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 715–718, 2010.

[201] Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access*, 8:176274–176285, 2020.

[202] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[203] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63, 2011.

[204] Chi Sun, Luyao Huang, and Xipeng Qiu. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[205] Ming Sun, Jufeng Yang, Kai Wang, and Hui Shen. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016.

[206] Pranav Suri and Nihar Ranjan Roy. Comparison between lda & nmf for event-detection from large text stream data. In *2017 3rd International Conference on Computational Intelligence & Communication Technology (CICT)*, pages 1–5. IEEE, 2017.

[207] Shaheen Syed and Marco Spruit. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*, pages 165–174. IEEE, 2017.

[208] Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324, 2007.

[209] Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*, 2020.

[210] Bill Thompson, Seán G Roberts, and Gary Lupyan. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038, 2020.

[211] Huu-anh Tran, Yuhang Guo, Ping Jian, Shumin Shi, and Heyan Huang. Improving parallel corpus quality for chinese-vietnamese statistical machine translation. *Journal of Beijing Institute of Technology*, 27(1), 2018.

[212] Abinash Tripathy, Ankit Agrawal, and Santanu Kumar Rath. Classification of sentiment reviews using n-gram machine learning approach. *Expert Systems with Applications*, 57:117–126, 2016.

[213] Martin Trow. From mass higher education to universal access: The american advantage. *Minerva*, pages 303–328, 1999.

[214] Eka Surya Usop, R Rizal Isnanto, and Retno Kusumaningrum. Part of speech features for sentiment classification based on latent dirichlet allocation. In *2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, pages 31–34. IEEE, 2017.

[215] Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. Cross-media learning for image sentiment analysis in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 308–317, 2017.

[216] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[217] Xiaojun Wan and Jianguo Xiao. Single document keyphrase extraction using neighborhood knowledge. In *AAAI*, volume 8, pages 855–860, 2008.

[218] Yun Wan and Qigang Gao. An ensemble sentiment classification system of twitter data for airline services analysis. In *2015 IEEE international conference on data mining workshop (ICDMW)*, pages 1318–1325. IEEE, 2015.

[219] Hao Wang and Jorge A Castanon. Sentiment expression via emoticons on social media. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2404–2408. IEEE, 2015.

[220] Haopeng Wang, M Shamim Hossain, Abdulmotaleb El Saddik, et al. Deep learning (dl)-enabled system for emotional big data. *IEEE Access*, 9:116073–116082, 2021.

[221] Kai Wang, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. Region attention networks for pose and occlusion robust facial expression recognition. *IEEE Transactions on Image Processing*, 29:4057–4069, 2020.

[222] Tianyi Wang, Ke Lu, Kam Pui Chow, and Qing Zhu. Covid-19 sensing: Negative sentiment analysis on social media in china via bert model. *Ieee Access*, 8:138162–138169, 2020.

[223] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, and Xin Li. An optimal svm-based text classification algorithm. In *2006 International Conference on Machine Learning and Cybernetics*, pages 1378–1381. IEEE, 2006.

[224] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[225] Zeerak Waseem, James Thorne, and Joachim Bingel. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online harassment*, pages 29–55. Springer, 2018.

[226] Gary M Weiss and Foster Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of artificial intelligence research*, 19:315–354, 2003.

[227] Finis Welch. Education in production. *Journal of political economy*, 78(1):35–59, 1970.

[228] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. Inducing a lexicon of abusive words–a feature-based approach. 2018.

[229] Anna Wierzbicka. *Emotions across languages and cultures: Diversity and universals*. Cambridge University Press, 1999.

[230] Liang Wu, Fred Morstatter, and Huan Liu. Slangsd: building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Language Resources and Evaluation*, 52:839–852, 2018.

[231] Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, pages 1–51, 2019.

[232] Hui Yin, Shuiqiao Yang, and Jianxin Li. Detecting topic and sentiment dynamics due to covid-19 pandemic using social media. In *Advanced Data Mining and Applications: 16th International Conference, ADMA 2020, Foshan, China, November 12–14, 2020, Proceedings 16*, pages 610–623. Springer, 2020.

[233] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 29, 2015.

[234] Adams Wei Yu, Hongrae Lee, and Quoc Le. Learning to skim text. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1880–1890, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[235] Shorouq Zahra. Targeted topic modeling for levantine arabic, 2020.

[236] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics.

[237] Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. Machine translation of arabic dialects. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 49–59, 2012.

[238] Renbo Zhao and Vincent YF Tan. Online nonnegative matrix factorization with outliers. *IEEE Transactions on Signal Processing*, 65(3):555–570, 2016.

[239] Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3510–3519, 2021.