Towards Subjective Multimedia Summarization Framework for Sporting Event in the Context of Digital Twins

by

Samah Bader Aloufi

Thesis submitted in partial fulfillment of the requirements
For the Doctorate in Philosophy degree in
Computer Science

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

Abstract

Real-world events generate a massive amount of traffic on social media with live, moment-to-moment accounts as any given situation unfolds. The data generated on social media is loaded with the sentiment, opinions, and reactions of the public towards the events. Browsing the event-related data in its raw form would be an overwhelming task due to the extensive amount of data, making the search for any specific updates or useful information an especially daunting and time-consuming endeavor. Event-related data, if effectively summarized, could generate a comprehensive overview of the event in terms of what happened and how people reacted at the time. Unfortunately, most of the event-based summarization systems concentrate their effort on simply describing what happened during the event; ignoring a valuable resource of emotional reactions that portrays the event from varying perspectives.

Accordingly, in this thesis, we introduce an event-based summarization approach that incorporates multimedia data, sentiment, and human reactions with respect to their point of view in order to generate a concise subjective multimedia summary of the event. In order to achieve our goal, we introduce popularity prediction and sentiment analysis models, both of which are essential in our summarization approach. As the event unfolds over time, we utilize the popularity prediction model to extract a representative set of images for the event to be included in the summary based on the predicted popularity scores. Multi-modality features are exploited to develop our model, including various levels of visual features, textual features, and contextual features. In order to track users' opinions and changes in their sentiment in correlation with the occurrence of sub-events, we develop a sentiment classification model that effectively recognizes the sentiment conveyed in sport conversations. Due to the lack of a manually annotated sentiment dataset, we propose a new, manually labeled football-based sentiment dataset. We also create an automatically generated sentiment-lexicon specifically for the football domain. We assess the performance of our developed models and generated summary through extensive experimental evaluation. We explore the impact of different features on the performance of the popularity prediction and sentiment models. We also leverage the knowledge of sport fans to evaluate the generated subjective summarization through an online user-based survey. The experiment results confirm the effectiveness of our proposed solution for event-based summarization by considering multimedia data, sentiment, and people's subjective views of events.

Acknowledgements

First and foremost, no alphabet can tailor a thread of words that could embody my heartfelt gratitude and love for my dad: Bader. Father, you are a prominent source of love, strength, support, and everything to me in this life. What I have achieved, and will achieve in this life is possible because of you and your presence, your support, and your encouragement through everything, whether it be the best of moments or the most difficult. I must also express my deepest gratitude to my beloved mom: Souad, who is, unfailingly, the foundation of infinite love and support. There are no words that adequately articulate my appreciation to my parents for their many years of patience and sacrifice. A very special thank you also goes to my sisters and brothers for their support and encouragement. Their faith in me always motivates and inspires me in pursuing my dream. Thank you, my family, for being by my side and infinitely supporting and loving me.

I would like to extend my sincere gratitude to my supervisor, Prof. Abdulmotaleb El Saddik, for his invaluable guidance, advice, insightful feedback, patience, and understanding throughout my Ph.D study. Despite his busy schedule, he always finds time to discuss my research, help to shape my ideas, and listen to my concerns. During the years of working under his supervision, I gained considerable knowledge from him in relation to research, academia, and, in turn, life. Discussions with prof. El Saddik never fail to enrich my experience, widening my knowledge and insight. Without the endless support and assistance of prof. El Saddik, this thesis would not have been possible.

Additionally, I am grateful to Dr.Shiai Zhu, for his valuable feedback and assistance, his encouragement and inspiration, and hours of discussions and meetings. I am fortunate to have learned from his expertise and scientific knowledge. Thank you, Dr.Shiai. I am thankful to my friends for their help, support and encouragement during the long hours of working together over this challenging time. I especially owe my heartfelt gratitude to Hawazin Badawi and Rajwa Alharthi for their help, long hours of discussion, and simply being a great friend.

Finally, I would like to thank all of my Multimedia and Communications Research Laboratory (MCRLAB) colleagues for the good times we had together in the most friendly lab on campus, where we treat each other as a family. You are all amazing researchers, and I feel very fortunate to know and learn from each one of you.

Dedication

This dissertation is dedicated to:
My parents, Bader and Souad,
and my sisters and brothers.
You, my beloved family, who are always there for me, and have forever believed in me and my dreams, I thank you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Human beings are curious and inquisitive. They are eager to know the details of what is happening around them, near or far away, whether it be as serious as a natural disaster, contentious as a political debate, or as frivolous as professional sporting events. Traditional news media, including newspapers, television, radio, and newswire, has been the main source of information regarding worldwide affairs for decades. These traditional media outlets have provided partial coverage of events, often with a considerable amount of delay caused by the availability of reporters, the time it takes to arrive at the site of the event (in the case of an emergency), and the production process leading up to the actual airing of the news report [144]. The development of social media has transformed the way information is generated, disseminated, and consumed [100]. Social media has become a powerful hub for sensing real-world events as millions of active users are producing and propagating information related to a wide spectrum of events as they happen, making it the most valuable resource for real-time access to information in relation to global and local happenings. Significant real-world occurrences are chronicled on social media daily. For example, Twitter was flooded with reports and public reactions in regards to the explosion during Boston marathon in 2013 [118]. Real-time User Generated Content (UGC) during events contains a vast and valuable pool of human opinions, feelings, and reactions towards various events, as oppose to traditional media, which provides simply an objective and cursory view of the events. Consequently, social media is evolving into an invaluable tool for monitoring and tracking events in real-time which presents infinite opportunities to benefit a wide range of applications and society sectors.

The exponential growth of social media has contributed to the generation of a tremendous amount of event-related content, which raises challenges in tracking specific event evolution and identifying the major sub-events in relation to a specific main event. Sub-events are the key and notable moments of interest which occur over the course of a particular event. For example, during a football match, goals and red cards would be considered sub-events. Searching for popular event updates through social media can result in hundreds of thousands of search results, which can be overwhelming for users to encounter each time they want to check for updates. In addition, the large amount of duplicated, noisy, and irrelevant information that users may come upon can make finding pertinent information about the major sub-events that transpired during the event an increasingly difficult task. As the event progresses, people wish to know the updates from the events instantly and with less effort. This calls for techniques to efficiently monitor, analyze, and summarize the generated data during the event time span. Therefore, an effective event-based summarization mechanism that covers the critical moments of events is a necessity.

Recently, researchers have capitalized on social media's provision of a large volume of easily accessible and valuable real-time data in their studies involving event summarization. Traditionally, event summarization is defined as "creating a summary of events based on bursty features identified from a text corpus" [36]. To date, researchers have focused solely on textual data when mining social media information for event monitoring and summarization. Indeed, event related data shared on social media is not in the form of textual data only; images and videos are common types of shared media in capturing scenes before, during, and after the event. Exploiting multimedia data for event representation and summarization will be of a great benefit as it provides a vivid overview of the event, taking into accounts the fluid emotional states, opinions and reactions of users. Recent existing works, such as [121, 13, 111, 50], summarize events considering the rich multimedia event-related data available on social media. The main idea behind previous studies in terms of event summarization is to provide effective concise summaries that answer when, where, and what happened during the event's lifetime. This type of summarization presents only the objective view of the event based on factual information, while the reactions and the opinions of the public toward the event are undervalued and ignored. Including the diverse human opinions and reactions towards real-world situations in event summarization will provide meaningful insight into understanding the key aspects being discussed during the events [29]. This information can be easily accessed through social media where people post their opinions, reactions, and views of the event as the event develops, making it a promising and limitless tool for study. In this thesis, we focus on

monitoring and summarizing human reactions and feelings during the events utilizing multimedia data. Our aim is to go beyond the mere objective description of the event. What if we could provide the public with the digital twin of the event? What if we could let the public experience the event through a digital replica of the event that encompasses the meaningful summary of the event happenings, and carries the intensity emanating from the human feelings and reactions lived during the event? In order to achieve this goal, we need a model that can converge multimedia technologies and enable monitoring and interpreting various types of data to provide a comprehensive subjective summary that demonstrates the public reactions towards event happenings.

Digital Twins is a compelling technology of recent fruition, which aims to build a digital replica of a real-world entity that can continuously monitor and detect changes in the real twin status, as well as trigger an action when an anomaly pattern occurs. Digital Twins, as envisioned by El Saddik [40], are defined as "digital replications of living as well as nonliving entities that enable data to be seamlessly transmitted between the physical and virtual worlds." The digital twin is the convergence of multimedia technologies that facilitates monitoring, understanding and optimizing the functions of all physical entities, living and nonliving [40, 41]. In a nutshell, the Digital Twins technology automatically monitors the real twin and analyzes the collected data to build a replica of the real twin. Thus, we propose a digital twin based approach that captures people's reactions to the event happenings, analyze it, and summarize the real-world event data to provide a comprehensive overview of the happenings during the event and illustrate the public reactions towards the event based on their point of view. Specifically, we concentrate on generating multi-modal subjective summarization for sport-based events.

Individuals experience events and situations differently based on their perspectives. Sporting events, especially, produce a wide range of emotions and reactions that are shared and discussed by millions of spectators worldwide. Fans describe the event in various ways based not only on their expectations and experiences [138]; but also, the strong feelings evoked throughout the span of the event. Most sport fan posts are completely subjective and express personal opinions about the event based on their point of view, making sporting events very conducive to subjective summarizations that illustrate two distinct views of the same event, in this case, centered on which team they are supporting. Football, or soccer, is the most popular among all professional sports, with 4.0 billion fans all around the world[1]. Football events such as the FIFA World Cup and the UEFA Champions League attract millions of fans worldwide. More than 3 million people attended games

[1]https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html

in the FIFA World Cup 2018 tournament, and 3.6 billion people watched the event live online or on television. Twitter is the most popular platform for football fans to post real-time messages during football events, sharing their experiences and opinions. In the 2016 UEFA European Championship, 109 million tweets were sent by fans related to the tournament, with 14 million of them referring to Portugal's defeat of France in the final match alone[2]. According to Twitter's official blog, 115 billion tweets were recorded during the FIFA World Cup 2018 as people reacted to what occurred during the tournament such as goal scoring, players injuries, or prediction of the next match outcome[3]. This illustrates the fact that social media, namely, Twitter, has become the go-to venue for football fans to discuss and express opinions and emotions regarding the sub-events that happen during matches with respect to their team and opponents. Accordingly, we utilize football-based data to automatically generate a subjective multimedia summarization that sheds the light on fans reactions and changes in their feelings during any given football match.

1.2 Motivation

Sporting events incite a wide range of emotions and reactions that are shared and discussed by millions of spectators worldwide, generating a vast amount of data on social media. Often, this data includes moment-to-moment accounts of the events. This massive volume of opinionated social stream reflects people's opinions, sentiments, views, and emotional reactions in regard to what is happening during the event, according to each individual's subjective point of view. The combination of a wide range of personal perspectives and live coverage makes social media a significant venue for monitoring the events, and understanding their impact on the public, while covering the different aspects discussed during the course of the events. These diverse subjective reactions and the intensified emotions of fans, if summarized and analyzed, provides an insight into how people perceive the event, and, can in turn, aid in predicting possible consequences. Various stakeholders can capitalize in different ways from summarizing and analyzing sporting events in real time. Sports organizations and websites can provide their customers with timely updates on key events occurring during the matches and, at the same time, highlight significant reactions of fans from both teams. Analyzing fans' opinions and reactions will help to illustrate the change in their sentiment in alignment with the occurrence of sub-events. Fans can benefit from

[2]https://blog.twitter.com/official/en_us/a/2016/fans-turned-to-twitter-as-portugal-won-euro2016.html

[3]https://blog.twitter.com/en_us/topics/events/2018/2018-World-Cup-Insights.html

such a subjective summary by comparing their emotions and reactions during the historical moments of the event to those of other spectators. City municipalities and authorities can use these real-time summaries of sporting events to monitor the change in fans' sentiment, flagging intense negative turns in reactions, and then take precautions in anticipation of possible unrest and violent incidents like riots. For example, during the FIFA World Cup 2018 competition, Argentina fans started a fight against Croatia fans after the match where Croatia defeated Argentina 3-0. This fight was recorded and circulated on social media, leading the Argentinian authorities to take action, requesting the deportation of the fans who participated in this violent action[4]. Moreover, the match between Brazil and Serbia stirred up negative and angry emotions among the fans when Brazil advanced to the round of 16 stage, eliminating Serbia. The violence after the match resulted in the arrest of nine Brazilians by the police due to their violent behavior[5]. Conclusively, effective summarization which provides a meaningful description of an event's significant moments with respect to fan perspectives and feelings could prove useful in a variety of applications. Thus, in this thesis, we aim to produce an automatic summary of a football event that summarizes the match by exploiting textual and visual social data posted by users during important moments throughout the event, including the subjective point of view of the fans and their reactions as the event unfolds over time.

1.2.1 Motivating Scenario

In this section, we provide scenarios that highlight the advantages of generating effective subjective multimedia event-based summarizations.

Let us take John as an example. John is a full-time professor who manages a research lab with many graduate students, besides his duties as a father. John is also a football enthusiast. His work and family responsibilities prevent him more often from watching his favorite team's games as often as he would like. At this point, John is looking for real-time updates about new events which occur during matches such as goals, penalties, or red cards with minimal time and effort spent reading and browsing through the status updates on social media. Thus, to help John find out the information directly with less effort, a summarization system would provide him with prompt and specific information regarding the event. Would a text-based summarization be enough to provide a complete

[4]https://www.telegraph.co.uk/news/2018/06/22/croatia-fan-beat-world-cup-match-argentina/

[5]https://www.standard.co.uk/sport/football/worldcup/world-cup-2018-dramatic-moment-serbia-and-brazil-fans-clash-in-the-stands-in-russia-a3874326.html

view of the event for John? The answer is no, as, what if John wanted to see the popular pictures posted depicting the goal scoring moment?

Let us consider another example: during the match, there were a few incidents that occurred which got John fired up. He would like to commiserate with other fans, sharing opinions and reactions, in real time. He would also like to see comparisons of how his team's fans perceive the situations during the match in comparison to the reactions of the opposite team fans. John is a busy researcher who has little time to browse a huge amount of data in order to explore how fans of both teams see and feel about the events. The availability of event-based data on social media that includes opinions and reactions of fans regarding situations occurring during the match time make the development of our subjective summarization that leverage multimedia data a possible solution for this problem. Consequently, our proposal for automatically generating multimedia event summarizations that present the two views of the same event based on the fan perspectives in real-time could be very useful in aforementioned scenarios.

1.3 Research Objective

In order for a sport-based event summarization system to truly capture a vivid and realistic picture of a particular event, it must consider all data available regarding the occasion, especially the subjective content that drives directly from those who possess the most emotionally vested interest- the fans. Previous studies have only considered the single modality, or objective, factual view, and, in so doing, overlooked the considerable and useful multimedia data that describes not only each stage of the event as it develops in real time; but also, the subjective data which illustrates fan emotions and reactions related to the event as it evolves. Our objective in this thesis is to take advantage of this wealth of introspective information in order to automatically generate a subjective multimedia sport-based event summarization that pinpoints and outlines the different impressions of the same event based on users' varying point of views. Our research is divided into three main stages. In the first stage, we develop a popularity prediction model to rank social images based on their predicted popularity scores. Since we are summarizing a fresh, real-time event, it is essential to predict which images will be popular at the time in order to extract a sub-set of visual data that appropriately represents the ongoing event and its fans' reactions. In the second stage, we focus on developing a sentiment classification model to track the change in fans' sentiment throughout the event. The classifier categories each fan tweet as positive, negative, or neutral in order to create a blueprint of fan sentiment with respect to the

occurrence of sub-events during the match. In the final stage, our summarization pipeline is proposed, which leverages the developed models to produce subjective multimedia event-based summarizations of important moments which occur during an event. Our proposed solution is not limited to answering simply what happened during the event; it has the ability to summarize how people feel and react to what transpires as well.

In this thesis, we address a series of conditions that are related to summarizing a live sporting event in real time by undertaking the following steps in producing a subjective multimedia summarization:

- Support the prediction of popular images during the event which will be included in the summarization as a visual representation of the happenings during the event. This requirement is addressed in chapter 4.
- Support the analysis of fans' sentiment in correlation with the occurrence of important sub-events as the event develops over time. This requirement is addressed in chapter 5.
- Detect the occurrence of significant junctures during the event and assign the generated event-based data into team sets. This requirement is addressed in chapter 6.
- Summarize the event utilizing both visual and textual modalities which illustrate the differing views of the same event. This requirement is addressed in chapter 6.

This thesis also contributes in addressing the following research questions:

- Can we predict which image will be more popular than others on social media by studying the effect of various factors?
- How to effectively classify sentiment appearing in football-specific conversations?
- How fans react to what happens during an event?

1.4 Thesis Contributions

The objective of this thesis is to summarize a sporting event based on the subjective views of its fans, utilizing the multimedia data that are available on social media. The main contributions of this thesis are as follows:

1. Proposing a subjective multimedia sport-based event summarization framework based on the Digital Twin technology which not only generates a chronological summary of the happenings that occurred during the event, but also considers the differing views of the public towards the same event while exploring sentiment changes among users over the course of the event. We also integrate the visual aspects of the event in our summary in order to provide users with the most popular images posted during the event.

2. Designing and implementing a sub-event recognition algorithm which is able to detect the occurrence of significant moments within the main event by monitoring the increase in volume of the social stream. The algorithm can differentiate if the increase in the social interactions belongs to similar sub-events or signals a new incident, and then labels the new occurrence using the most representative key terms.

3. Modeling and implementing the tweets categorization algorithm that identifies if a tweet is written by a user from a specific team's fan base. We propose a two-stage approach for assigning tweet collection to a specific team by utilizing the hashtags and keywords which appear within the tweets, and leveraging the user's social profile to corroborate their preference for a specific team.

4. Designing and developing an image type classification model to classify a given image by its type. The developed model can distinguish between different synthetic image classes and a natural image class, and is employed to filter out inappropriate images for our summarization purposes.

5. Designing and developing a multi-modality popularity prediction model to predict the popularity score of a given image. In the process of developing our model, we experiment with combining various factors and investigate their effect on predicting a given image's popularity score. Specifically, visual features, including low level, middle level, and semantic features, as well as textual and contextual features are investigated in providing a comprehensive understanding of images popularity prediction.

6. Constructing a Football-specific sentiment dataset which includes tweets collected from two football events: the UEFA Champions League 2016/2017, and the FIFA world Cup 2014. Each tweet in the collection is annotated manually by four annotators, and labeled as positive, negative or neutral tweets. Our dataset consists of 54,526 labeled tweets.

7. Developing a new sentiment lexicon utilizing our constructed Football-specific dataset. The Football-sentiment lexicon is generated automatically utilizing a corpus-based approach in order to facilitate the development of a domain-specific sentiment lexicon and improve the performance of the sentiment classifier.

8. Designing and developing a Football-specific sentiment analysis model capable of recognizing the sentiment expressed in football related conversations.

9. Providing a comprehensive experimental evaluation to measure the effectiveness of our subjective multimedia summarization approach. Each main component in our approach is evaluated in terms of accuracy and effectiveness on real world data.

1.5 Scholarly Achievements

- **Research Resulted in Refereed Journals**

  1. **Samah Aloufi**, and Abdulmotaleb El Saddik. "Sporting Events Digital Twins: A Multi-view Multi-modality Framework for Sporting Events Summarization", Submitted.
  2. **Samah Aloufi**, and Abdulmotaleb El Saddik. "Sentiment Identification in Football-Specific Tweets" IEEE ACCESS 6, pp.78609-78621, 2018.
  3. **Samah Aloufi**, Shiai Zhu, and Abdulmotaleb El Saddik. "On the Prediction of Flickr Image Popularity by Analyzing Heterogeneous Social Sensory Data." Sensors 17, no. 3 (2017): 631.
  4. Shiai Zhu, **Samah Aloufi**, Jun Yang, and Abdulmotaleb El Saddik. "On the Learning of Image Social Relevance from Heterogeneous Social Network." Neurocomputing 210, pp.269-282, 2016.

- **Research Resulted in Refereed Conferences**

  1. Rana Abaalkhail, Fatimah Alzamzami, **Samah Aloufi**, Rajwa Alharthi, and Abdulmotaleb El Saddi. "Affectional Ontology and Multimedia Dataset for Sentiment Analysis". In International Conference on Smart Multimedia, 2018.
  2. **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). IEEE, 2018.

3. Shiai Zhu, **Samah Aloufi**, and Abdulmotaleb El Saddik. "Utilizing Image Social Clues for Automated Image Tagging." In Multimedia and Expo (ICME), 2015 IEEE International Conference on, pp. 1-6. IEEE, 2015.

1.6 Organization of the Thesis

The remainder of this thesis is presented in chapters 2 to 7. The content of these chapters is summarized as follows:

- Chapter 2: introduces existing works and background research that are related to problems addressed in this thesis. It reviews related works that were conducted in the area of social image popularity prediction, as well as summarizing sporting events on social media. This chapter also discusses various studies that belong to different sub-categories of event summarization, including text-based and multimedia-based summarization. Since our focus is sentiment analysis of football conversation on social media with the goal of summarizing football events, this chapter also details the existing works that contain relevant information to this aspect of our work.

- Chapter 3: presents an overview of our designed framework for generating subjective multimedia summarization, and highlights the importance of each component. A brief description of each component's task is provided in this chapter, while a comprehensive explanation of each component's development process is given in the following chapters.

- Chapter 4: discusses the design and development process of the social image popularity prediction model. The chapter first provides a comprehensive analysis of social interaction on social media and describes the factors that could impact the social popularity. Then, a detailed description of the development process of the popularity prediction model is presented. It explains, in details, the multi-modality features including visual, textual, contextual features that are used for model training. Also, a comprehensive evaluation of the performance of individual features is described. Various fusion techniques are utilized to investigate the influence of combining different features on the model performance and the results are reported in this chapter.

- Chapter 5: explains the process of developing the sentiment classification model, which is specifically designed for the football-specific domain. It starts by describing the construction of our football-specific sentiment corpus. It then provides the details

of building our sentiment model, including data preprocessing, feature extractions, and learning algorithms. This chapter also discusses the performance of different learning algorithms in classifying the sentiment conveyed in football tweets, as well as the impact of the utilized features on the effectiveness of the sentiment model.

- Chapter 6: introduces the details of the subjective multimedia event-based summarization approach, which include the detection of important moments within a stream of social event data, and the identification of the type of the occurred sub-event. Then, it discusses the next step in the approach, which is to assign each tweet to a team. Next, a description is given of the process of extracting representative sets of textual and visual data in order to summarize the event. Finally, it provides the experiment results of evaluating the effectiveness of the generated subjective multimedia summary.

- Chapter 7: summarizes the thesis content and contributions, and briefly discusses possible future research directions.

Chapter 2

Background and Related Works

In this chapter, we discuss the most relevant literature to the research problems encountered in this thesis. Particularly, we study resources that use various approaches for predicting images popularity on social media. We also examine related works which focus on studies that specifically address football sentiment analysis. Finally, we provide detailed description of existing studies that discuss event summarization on social media.

2.1 Images Popularity Prediction

Our objective is to understand the social interactions among different entities on social media, and to predict image popularity. Thus, we group previous works into two categories: (1) human social behavior on the Web; and (2) image interestingness, or attraction, and popularity.

2.1.1 Social Interactions between Users and Data

Multiple existing studies focus on the analysis of social behaviors and information propagation on social media. In [139], Van Zwol studied the characteristics of social behavior in users on Flickr. This work demonstrated that photos received the majority of their exposure within the first two days of being uploaded; furthermore, these views were influenced by the account owners' contacts and social groups to which the owner belonged. Van Zwol also documented that images with a high number of views are explored by users worldwide. In comparison, Cha et al. [27] analyzed the propagation of photos on Flickr and also noted that a user's social links had a significant impact on image popularity and

proliferation. Unlike [139], however, they evaluated the favoritism that images received, and found that these were coming from users within a few clicks of the owner or their network. The difference being that when considering only the number of views, like Van Zwol [139], since viewers are not required to be Filckr members, the expanse of a particular image's propagation in terms of its views alone can be more widespread. On the other hand, when considering the favorite action, which is only available to registered members, the reach is generally influenced by social relations. So, while we note that these two different studies drew similar conclusions in regards to a user's social links having a considerable impact on image appeal and distribution, their reference, in comparison, serve to highlight the fact that the various platforms of social media provide a multi-faceted resource in the examination of social behaviors and image propagation. A recent study conducted by Lipczak et al. [84] further scrutinized the social behavior of users who respond to photo posts by marking them as favorites on Flickr. They found that 50% of the responses occurred within one week of an image's upload date, and the largest portion of these favorites was from the owner's contacts. Alves et al., in [3], also studied the impact of users' networks on favorite actions on Flickr. They reported that 70% of the favorite indications on photos were received from the owner's contacts. Lerman and Jones [81] explored social behaviors on Flickr utilizing sample images from the Explore page of Flickr, a popular group called Apex group, and a random set of images. They analyzed the visibility of the images to users by considering the size of a user's network, the number of groups to which they belong, and the tags attached to the images. Their analysis showed a high correlation between an image's popularity signals - "views, favorites, comments"- and the number of reverse contacts of the owners, while tags were deemed less important. Social groups only interrelated with the random images' popularity and showed no essential role in Apex and Explore images. In Cha et al.'s work [26], they analyzed favorite behavior on Flickr and identified two patterns: the number of favorites for an image usually increases steadily, but certain images, at times, experience rapid growth because of external exposure. On the other hand, Valafar et al. [136] also studied the favorite behavior on Flickr and reported that 10% of users are the cause of 80% to 90% of the favorite action. Additionally, they confirmed the findings of other studies asserting that photos are generally discovered within the first week of being posted, after which the popularity will increase steadily during the photo post's lifetime. We can see that analyzing human social behavior is a challenging and multi-dimensional task; thus, a stable and unified pattern is difficult to model. This is in contrast to other media types or platforms. For instance, YouTube videos are constantly viewed by users throughout their time online; whereas, news often reaches a saturation point within a few hours of being posted [130].

2.1.2 Image Interestingness and Popularity

Defining image interestingness or, its appeal, is difficult because this subjective concept is entirely based on a user's preferences. Despite the difficulty, we can still observe agreement among a large number of online users on image attractiveness, especially with the proliferation of social media. This phenomenon attracts researchers from computer vision and multimedia domains who attempt to identify the driving forces behind the fact that certain photos are considered popular or interesting. In addition to the photo's content itself, the studies conducted used either a single modal, or a multi-modal approach which combines multiple features that are available within the social networks in order to predict an image's popularity. Because users on social media sites can view, comment and select as a favorite any image to express their interest, these social metrics can be utilized to indicate a photo's popularity [27]. While utilizing different social metrics to predict popularity, researchers have modeled the problem of popularity prediction with several learning paradigms.

Van Zwol et al. [140] modeled the prediction of whether a given user will select an image as a favorite as a binary classification using Gradient-Boosted Decision Trees (GBDT). The learning algorithm was trained based on visual, social and tag features. They reported that, in most cases, combining social and visual features achieved the best results in predicting the image that a user will mark as their favorite. McParlane et al. [93] also addressed popularity prediction as a binary classification by training a non-linear SVM classifier. They considered the cold start scenario, where interaction information is not available. In the cold start case, there is limited textual and interaction information; therefore, information related to image and user contexts was utilized. They used the number of views and comments as the two social popularity metrics to classify images into low- or high-popularity classes. Each image was represented as a binary vector according to the extracted visual feature, social clues and textual information. The results showed that a user's social context played a significant role in predicting image popularity.

Meanwhile, Hsieh et al. [54] found a weak correlation between an image's aesthetics and its popularity on social media. They attempted to determine which image features led to social popularity and concluded that, among basic image features, color was the most influential. Their work verified that the beauty of a photo does not guarantee its popularity on social media. The idea of addressing the popularity problem as classification and regression instances was introduced in [120] by San Pedro and Siersdorfer, where classification models were used to sort photos into attractive or unattractive classes, and a regression model was used to rank a photo based on its attractiveness. They combined

basic visual features and tags to predict image appeal defined as the number of favorites. As expected, the results showed that combining tags and visual features performed better than relying on only one feature. As these studies only considered part of the factors or representations in social data; a more comprehensive study is needed.

Khosla et al. proposed more comprehensive work in image popularity prediction [70]. They built a regression model to predict image popularity based on the number of views normalized by time. They investigated the effectiveness of low-level visual features, such as texture and Gist, and as well as that of deep learning features. In addition to the content features, they leveraged the social clues related to the users and images; for instance, contacts, groups, tags and the lengths of titles and descriptions. When combined, visual and social features led to a better performance than individual features. Following a similar approach as [70], in [43], Gelli et al. utilized visual sentiment features and high level features (objects). They also employed social and textual features to recognize named entities within the text attached to the images, referring to them as tag type and tag domain.

Yamasaki et al., in [149], proposed to predict image popularity scores by examining tag information. They computed the tag's importance by combining the tag frequency and weight learned through support vector regression. This method obtained better results than predicting popularity based solely on tag frequency scores. Their approach is cost effective, but is reliant on the presence of tag information in order to predict the popularity score; therefore, if an image has no tag information, this method is not applicable. Learning the rank of an image, in terms of popularity based on visual features, was introduced in [25] by Cappallo et al. They used deep learning features to discover latent scenes that affect the popularity of images. Another approach was proposed by Wu et al. [148], who adopted a matrix factorization technique to predict image popularity, using temporal information, specifically, the interaction between users and images.

Prior works considered various types of features related to image content (content based) and/or social features. These features helped predict a social image's popularity by addressing the problem as a learning problem with different paradigms. There is no unified definition of image popularity or of evaluation metrics due to the difference in the way the problem is approached by different groups. As of the time of this dissertation's publication, the most recent work on image popularity prediction was conducted by Hsu et al. [56]. They integrated multi-modal features in order to capture the relationship between an image and its popularity. They used numerical data (user id, post id, geographical information, and the date of the post), textual features, and categorical data along with the visual

content of the image for learning the popularity prediction model. In another recent work, Kang et al. [68] proposed a Catboost-based framework for image popularity prediction. In their work, they used features related to the posted images such as, title, geographical information, and the date the image was posted. Besides image features, they employed user information to train a regression model for the popularity prediction task.

2.2 Football-based Sentiment Analysis

It is estimated that, currently, over 3 billion people worldwide use social media, making it an invaluable resource with infinite applications[1]. While the percentage of the world population using social media rises yearly, sentiment analysis, as a field of study, is seeing a surge in interest and function. Sentiment analysis or opinion mining is defined as the computational study of opinion, sentiment and emotions expressed in text towards a specific entity [85]. Sentiment analysis has been applied to several types of text including movie reviews, news and blogs, product reviews, and tweets. A considerable amount of research focuses on analyzing sentiment expressed in tweets including Balahur in [8], Giachanou et al. [44], Go et al. [47], Kouloumpis et al. [74], and Pak and Paroubek [107].

Few studies, however, are conducted using football-specific sentiment analysis. Barnaghi et al. [9] utilized a logistic regression algorithm to learn polarity (positive and negative) classifier based on Uni-gram and Bi-gram features. They achieved an accuracy of 72% using the Uni-gram feature. In a similar manner, authors in [10] combined N-gram features with external lexicon-based features to improve the performance of the Bayesian logistic regression classifier. The best performance of their proposed method was obtained through combining Uni-gram and Bi-gram features, which resulted in an accuracy of 74%. Both [9] and [10] first used 4,162 manually labeled tweets to build their sentiment model. They then collected tweets during the FIFA World Cup 2014 where they utilized the learned sentiment classifier in order to find correlation between sentiment and major events which occurred during the competition. In [42], Alves et al. proposed the sentiment analysis method for football related tweets written in Portuguese. In order to train their sentiment models, they collected tweets during the 2013 FIFA Confederations Cup. The tweets were labeled based on two methods: automatically based on the positive and negative emoticons included in the tweets, and, a random sample of 1,500 tweets which were manually labeled. The best performance accuracy was achieved by SVM (87%) when trained and tested on

[1]https://blog.hootsuite.com/social-media-statistics-for-social-media-managers/#general

the automatically labeled data. In contrast, this accuracy dropped to 66% when trained and tested on the manually labeled dataset.

In another case, Gratch et al., in [49], used lexicon-based features to train the Naïve Bayes algorithm on the SemEval 2014 dataset [116]. They approached the problem of sentiment analysis by classifying tweets into positive, negative or neutral classes. They proposed training the sentiment classifier on a manually-labeled general dataset, then using the model to identify sentiment in football related tweets. To evaluate the sentiment model performance on football data, they manually labeled 154 tweets related to the FIFA World Cup 2014. Their results showed a strong correlation between the ground truth and the algorithm results; however, the validation set consists of such a small number of tweets, it may not reflect the overall performance of the sentiment model on football related tweets. Aloufi et al., in [2], trained the SVM classifier on the FIFA World Cup 2014 dataset utilizing N-gram features and various lexicon-based features. The dataset used for training in [2] is automatically labeled, which makes it vulnerable to incorrect labeling. Their proposed method achieved an accuracy of 85% in classifying tweets into positive, negative, or neutral classes.

Previous studies in football sentiment analysis followed machine learning and lexicon-based approaches, which are widely used in sentiment analysis. The sentiment lexicons utilized in previous works, such as [49] and [10], are general sentiment lexicons which are generated from diverse resources and domains. Sentiment analysis is dependent on the domain in which it is applied because a particular word could convey different sentiments and meaning in various domains [145, 78]. Furthermore, the choice of lexicons as features for sentiment classification is very important; as the classification performance could be compromised if chosen improperly. For instance, using the EmotionWord lexicon to train a sentiment classifier on sarcastic tweet sets actually jeopardizes its performance [89]. Using a combination of lexicons in [124] has shown to be very effective on the FIFA dataset. This improved performance on the FIFA dataset could be due to a high level of emotions and opinions contained in fans' tweets [49]. This shows the need for developing football-based sentiment lexicon.

The machine learning approach relies on a labeled dataset in order to train a classification model. Publicly available sentiment datasets can be divided into: general datasets such as [107], [47], or [99], and domain-specific datasets. The domain-specific datasets include the Health Care Reform dataset [127], the Sanders[2] dataset which consists of tweets from four different platforms: Apple, Microsoft, Google and Twitter, the Dialogue Earth

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

Twitter Corpus[3] which contains three datasets: WA and WB for weather, and GASP for gas price. For the football domain, the publicly available dataset designed specifically for football tweets sentiment analysis is the FIFA World Cup 2014. The FIFA World Cup 2014 dataset, introduced by [9] and [10], consists of 30 million tweets collected over the course of the game. Each tweet was automatically labeled as either positive, negative or neutral, using Aylien Text analysis API[4]. The FIFA 2014 dataset is the first sentiment dataset designed specifically for football events; however, it is automatically labeled utilizing general sentiment algorithm, and, as such, it is our position that we cannot rely on an automatically labeled dataset for accurate results due to the noisy labeling produced by this approach of annotation.

In order to overcome the limitation of previous works, we propose the development of a football-specific sentiment classifier that can effectively recognize sentiment in football related tweets. To do that, we propose the creation of a new and improved football dataset which is manually labeled to support the research in sentiment analysis for football tweets. In addition, we intend to develop a new sentiment lexicon using a corpus-based approach.

2.3 Social Event Summarization

In this section, we review some automatic summarization techniques that we consider important for this work. We focus on studies that address social media summarization. We have divided the existing works into three categories: text-based, multimedia-based, and sport-based summarization.

2.3.1 Text-based Event Summarization

In the early stages of automatic text summarization, the process focused on summarizing a single document. The summarization system would scan a document and then generates a summary of the important information embedded within the original document. As the web became more popular and presented as a powerful tool, researchers evolved their studies, endeavoring to summarize multiple related documents at once [119]. The multi-document summarization systems are grouped into two categories: abstractive summarization or extractive summarization. Abstractive summarization produces a new text that encapsulates the original document and may include new vocabulary while keeping the

[3] www.dialogueearth.org

[4] https://aylien.com/text-api/

same meaning. In contrast, the extractive summary is generated by extracting the most significant sentences from the original documents. Microblog (Tweets) summarization is considered a special case of multi document summarization.

Several studies have been conducted around text summarization on social media. Sharifi et al. in [123] used a graph-based algorithm to find the most commonly used phrases that appeared in a collection of tweets. Then, a sentence was selected to summarize the event, based on the set of the identified phrases. Authors in [59] summarized Twitter posts using a clustering algorithm. Based on the similarity measure, tweets were grouped into k clusters and then each cluster was summarized by extracting the highest scored tweets, which were graded based on the Hybrid *TF-IDF* weighting scheme. Chakrabarti et al. [28] designed a sub-events detection algorithm by modifying the Hidden Markove Model (HMM). In order for HMM to learn the vocabulary of the event, it was trained on tweets from previous similar events, specifically a number of American football matches. Notably, while the HMM model is applicable for detecting recurrent events, it falls short when it comes to detecting new sub-events. Louis and Newman, in [88], summarized business-related tweets by employing a concept-based clustering method that groups tweets into sub-topic clusters, and then ranking for summary generation. Chua and Asur [31] designed two topic models for detecting events on Twitter. They incorporated the temporal factor with the words in tweets for the event discovery task. Then, a set of tweets is selected to describe each detected event.

The aforementioned works focus on summarizing social events utilizing only textual information. Recognizing that images and videos are also essential components in the documentation and description of social events on social media, lately, more studies consider summarizing social events by leveraging multimedia resources. The following section discusses the visual-textual type of social event summarization which utilizes social data.

2.3.2 Multimedia-based Event Summarization

Several attempts have been made at producing a multimedia summarization for an event or topic on social media. McParlane et.al in [92], proposed a visual approach to summarizing events by selecting and ranking significant images in order to maximize the relevancy and diversity within the set of selected images. The top ranked images were then used to summarize the event. McParlane's work utilized only images to summarize an event. Similarly, in [4], Amato et al. proposed a multimedia summarization technique using a graph-based approach. They focused on generating a visual representation using a Flickr

dataset. Alternatively, Cai et al. [22], developed a probabilistic topic model which jointly model selected Twitter features to discover events or topics from a stream of social data. They considered text, image, location, timestamp, and hashtag for topic modeling. Each event is then illustrated by selecting a sub-set of representative images.

Bian et al., in [13], focused on generating a combined textual-visual summarization by introducing a generative topic model named CM-LDA for sub-topic detection. The CM-LDA model discovered sub-topics from a collection of both text and image belonging to a specific social event. The event is then represented by illustrating these sub-events using a set of representative text and images. Qian et al. [111] proposed a cluster-based approach that jointly utilized text, images, and users for sub-event discovery. Their intuition to consider the user factor is based on the assumption that different sub-events attract the attention of specific groups of users. After data preprocessing, they employed their proposed text-image-user co-cluster algorithm to discover sub-events within the event dataset. The event summarization is then generated by selecting representative texts and images among all the discovered sub-events. Both [22, 111] have considered location attributes for event summarization based on the assumption that users within the same location are interested in similar topics and events. Their assumption is valid for local events that attract attention of people within the same area.

Schinas et al. in [121] introduced the MGraph framework which utilized both topic modeling and a graph-based approach to summarize social events. MGraph used the Structural Clustering Algorithm for Network (SCAN) and then the DivRank algorithm to detect sub-topics. The DivRank, which is a graph-based ranking algorithm, ranked images based on their social popularity and relevance to the events, then, a sub-set of top ranked multimedia be utilized for generating multimedia summarization. Guo et al., in [50], proposed another graph-based framework for event summarization on social media that is based on detecting the multi-aspects of the event and tracking their evolution. In order to do that, they utilized text similarity for aspect or clue discovery, and then employed a temporal-based clustering approach for clue segmentation and evolution tracking. They generated a multimedia visualization of the events by selecting a representative set of text using a dominating set function, image selected using a cross-media mining algorithm which outputs a set of relevant and diverse images. Qian et al., in [110], considered user opinion and proposed a multi-modal, multi-view topic-opinion mining (MMTOM) model for social event analysis. This model, however, was designed for news article texts which are well-structured, unlike social media data. This particular output is not a holistic summary; instead, it displays a set of representative concepts and words which appeared in the event

descriptions and top-ranked images.

Conclusively, these studies have all focused on generating an objective-based summarization of events. None of them considered human opinions or sentiments when analyzing the events data. Analyzing user point of view regarding an event is important in drawing conclusions about their feelings around what is happening. Researchers in [60], presented a framework created to generate textual summaries of soccer events based on real-time sentiment analysis. They used a burst detection technique to detect the occurrence of a significant event. These events would then be classified into one of five fixed classes. Based on the sentiment analysis of the tweets covering each event, they assigned the event to a team, then generated the text summarization of the match highlights.

All previously discussed works exploited different approaches for event summarization, all of which depend on event historical data. Furthermore, the selection of representative sets of multimedia-data for generating the summary utilizes the social popularity factor as a ranking criteria. Also, the output of these works are objective summaries which describe the sub-events or topics that occurred during the events.

2.3.3 Sport-based Event Summarization

Multiple studies are devoted to utilizing social media streams as a source for sporting event detection and summarization. Previous approaches can be classified into two main classes: graph-based or rate-based methods. Nichols et al. [101], were one of the first to address the possibility of sport event summarization. Their method is classified as rate-based, where a sub-event is detected when the tweets volume exceeded a pre-defined threshold. In their work, the threshold computed offline for the entire match using basic statistics. They then applied the phrase graph approach for selecting representative tweets which describe the sub-events occurring within the match. Zubiaga and co-authors [158] investigated the problem of real-time summarization of sub-events occurring during football matches. They flagged a sub-event occurrence when the rate of the tweets stream was above 90% of the previously seen rates. After detecting a sub-event, they utilized the Kullback-Leibler Divergence method for term weighting to select tweets with the highest scores for summarization of the detected sub-events. Marcus and others in [91] introduced the "Twitinfo" system for events detection and visualization. Based on specific keywords, Twitinfo detects sub-events using a peak detection algorithm, and then provide a timeline-based event visualization. The Twitinfo system identifies the sentiment and the location distribution of the tweets.

In contrast to [101, 158], Tagawa and Shimada, in [131], were interested in producing an abstractive summarization, which focuses on producing a summary using new vocabulary, instead of an extractive type of summarization which is based on selecting the most representative tweets. To generate the summary, Tagawa and Shimada first detected sub-events using a burst detection method, then extracted the elements which appeared in the sub-event, such as actions and players names. They first estimated the order of each element and its importance, then, they generated the summary of the sub-events utilizing fixed game-phrases. Kubo et al. [76] tackled the problem of producing sport event summarization in real-time by exploiting active users on Twitter who are known as good reporters. To select these reporters, they compute the user scores by assigning higher scores to users who, more frequently, post explanatory tweets during previously detected sub-events within the event. Gillani et al. [46] introduced the Lexi-Temporal Clustering (LTC) technique for producing post-event summaries of soccer games. They identified the occurrence of sub-events based on a fixed threshold that is calculated using the mean and standard deviation over all the match data. They then generated the summary by selecting the most representative tweets utilizing the K-means clustering algorithm. Zhao et al. [156], targeting American football, detected events based on the tweet rate change during a time sliding window. The detected events were then identified as either match-related or noisy events using a lexicon-based approach.

The work of Hsieh and others [55], proposed a real time approach for detecting sub-events which occur during a football match. Their approach included two steps: utilizing their moving-threshold method to detect sub-events, and using the TF-IDF technique to determine the most representative keywords for each sub-event. Similarly, Jai-Andaloussi et al. [61] leveraged social media content in summarizing football videos. They utilized the burst detection method introduced in [55] and applied the TF method as a weighting schema to rank the keywords in each detected highlight. They analyzed the video of the match and referenced only the highlights detected using tweets when creating the summary. In another work, Jai-Andaloussi et al. [60] detected sub-events and summarized soccer events utilizing the tweets themselves. After detecting the sub-event content from tweet threads using the moving-threshold method, they identify the type of event using a supervised machine learning approach. They built a classifier that categorizes tweets into one of five classes (goal, red card, yellow card, foul and penalty). They then used the sentiment of the Twitter posts during the highlight time to associate the event to a team. For the summary generation, they opted to select sentence templates approach to produce a summary similar to the one found in the football recap article. Van Oorschot et al. [138] sorted the identified sub-events into five fixed classes (goal, own goal, red card,

yellow card and substitution) by employing a Support Vector Machine (SVM) classifier. They assigned sub-events to a specific team by identifying fans based on the assumption that users will mention their preferred team more often over different matches. Using this classification method for sub-event identification limits the type of events that can be recognized during the matches. Sub-events that appear in the matches but do not belong to one of the specifically identified classes will be ignored which, in some cases, could result in the exclusion of critical sub-events.

Authors in [94] used a graph-based approach to represent a sequence of tweets as graphs where the sub-events were detected by leveraging the graph degeneracy concept. Subsequently, tweets were selected to describe the sub-event by applying a weighting schema based on term weight within the graph. Authors in [95] also proposed a graph-based approach to detect the rapid change in the graph's link weight for sub-events identification. They then employed a submodular function in order to extract the most representative tweets to describe the selected sub-event. Koleejan et al. [73] aimed for a more objective summarization and proposed ruling out tweets that were biased towards one team or the other. They calculated the subjectivity score of a tweet based on the word distribution demonstrated by the fans of each particular team.

All aforementioned works focused on generating objective summarizations of soccer matches. Football matches, however, are not objective events; as they invoke strong emotions among fans [49], and the supporters of each team have their own point of view regarding everything that happens in and around the event. Accordingly, a new approach to summarization is proposed in order to address subjective summarization based on multi-views or subjectivity and perspective. To the best of our knowledge, only two previous studies address the problem of generating subjective summarization for soccer. Each of the two works tackle the problem in a different way utilizing distinct techniques. Van der Lee and others, in [137], proposed a "PASS" system, which generates a Dutch summary of football matches by employing a template-based approach. Match statistics were plugged into the system and then the system would produce a textual description for fans of each team. They used a natural language generation technique to tailor the summaries of the soccer matches to the audiences from each club. The method used was dependent on utilizing a template report from each club or team website. Alternatively, Corney et al., in [33], proposed subjective summarization utilizing fan discussions on social media. First, they identified the user's preferred team based on how many times the user mentions a particular club. They then used a topic detection algorithm to recognize the topic of the sub-event by considering sudden increases in word frequency. They utilized a cluster-

ing algorithm to group the most descriptive phrases representing particular topics. Their study included only two matches in the analysis: The FA Cup final matches for 2012 and 2013. Two of the authors manually evaluated the resulting summary to determine if each summary corresponded to the comments posted by the BBC commentator. Both [137] and [33] provided text-based summarizations of the matches. To date, none of the works that summarize football matches consider leveraging multimedia data in order to present highlights of the event.

Chapter 3

Subjective Multimedia Summarization Framework: An Overview

3.1 Introduction

Sporting events generate a massive amount of traffic on social media with live moment-to-moment accounts as any given situation unfolds. This data not only captures on-the-spot information, but also represents fans' opinions, sentiments, thoughts, views, and reactions, centering on a specific event. All of this reactionary data is derived from each individual's unique perspective. Analyzing and summarizing this data will generate a comprehensive overview of the event in terms of how the event evolves, and how fans react to and view the event based on their perspectives. Prior studies involving social event summarization such as [13], [121] and [111], focus primarily on generating an objective-view summary. Generally, the extent of these studies is limited to providing a succinct summary that answers simply when, where, and what happened during the event's lifetime. This type of summarization provides only an elementary, fact-based, objective view of the event, while the reactions and the opinions of the public towards the event are undervalued and ignored.

Without a doubt, individuals experience events and situations differently based on their personal perspectives. They have diverse opinions, reactions, and their descriptions of the events vary based on their subjective point of view. Examining people's reactions and opinions regarding what is happening during the event is of significant importance in understanding the impact of the event on the public, and, in turn, predicting possible

consequences. Accordingly, developing an automated summarization mechanism that provides a meaningful description of a particular event's significant moments with respect to individuals' perspectives and feelings would extend the dimension of a single-perspective summary into a valuable subjective summarization. The subjective summary will create a comprehensive overview of the event by describing the same event from different viewpoints, shedding light on subjective individual reactions and opinions towards the event. In this thesis, we propose a summarization framework capable of generating a subjective multimedia summary for sporting events in real-time utilizing the Digital Twins technology.

In order to monitor and summarize sporting events as seen by fans in real-time, several key requirements need to be addressed. First, the summarization approach should recognize the occurrence of crucial moments within a continuous stream of data. Also, determining which tweets belong to each team is an essential step in producing a subjective summary based on fan perspectives. A sentiment analysis method must be integrated to the summarization mechanism so we can capture the emotions of the fans over the course of the event. As we are summarizing a fresh event in real time, we need to select a set of images that best describe the event happenings and illustrate fan reactions. Accordingly, an image popularity prediction model is proposed to predict which images will be more popular than others during the event in order to summarize the event. Finally, the summarization method should extract a set of selected images and text to help visualize the event.

Considering the above-mentioned requisites, we propose a sporting-event Digital Twins-based summarization framework developed to generate a subjective multi-modality summary of a football match in real-time. The proposed framework would capture fans' reactions based on the event's happenings, analyze their feelings, and produce a multimedia summary to provide a comprehensive overview of the event as experienced and described by fans. The framework consists of a set of components in order to analyze and summarize a given football match. First, a sub-event detection mechanism is employed to recognize the occurrence of an interesting sub-event and identify its type. Then, using a tweet categorization approach, tweets generated within the detected sub-event time frame are assigned to a team based on the tweet writers' fandom. Meanwhile, a sentiment analysis component designed specifically for football domain analyzes and tracks the changes in fans' attitude during the match. Simultaneously, our popularity prediction model predicts the popularity ranking score of images posted while the event is running. The framework handles the noisy images by employing a filtering mechanism to remove images that are unsuitable for summarization purposes. Finally, it selects a representative set of tweets and

images, generating a multimedia summary that illustrates how fans of each team respond and react to what is happening during the event.

In this chapter, we introduce the proposed framework through an overview description of the main components. The details of the development process for the popularity prediction model and the sentiment analysis model are discussed in Chapters 4 and 5, respectively. The pipeline of the framework will be introduced and explored in Chapter 6, along with the particulars of the input data, and finally, generating a subjective multimedia summary.

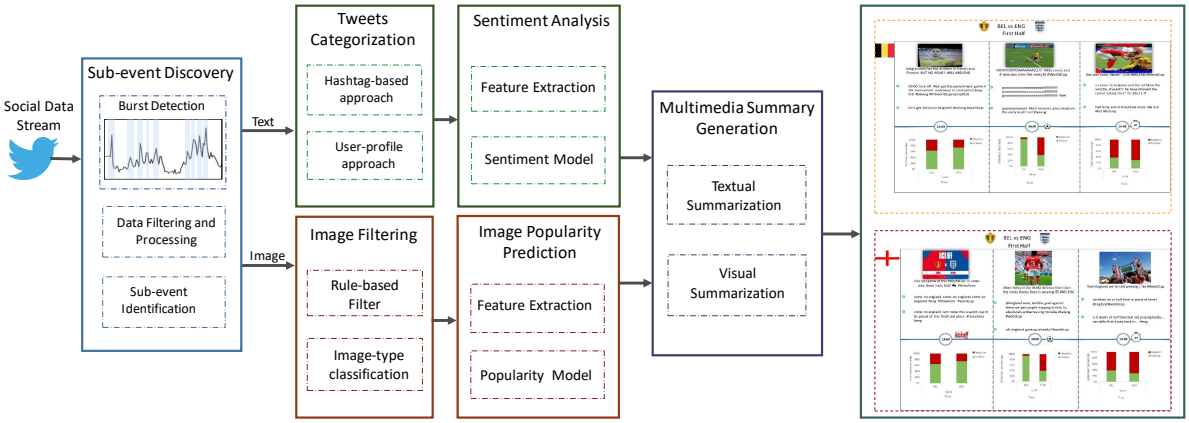


Figure 3.1: The proposed subjective multimedia summarization framework.

3.2 Framework Overview

Figure 3.1 illustrates the proposed subjective multimedia event summarization framework. The framework consists of the following main components:

- Social-stream Listener: We use the Twitter platform as a source for event relevant data collection. Twitter Streaming API[1] is used to collect microblogs related to a specific sports event in real-time.
- Sub-event Discovery: This component monitors the stream of social data and signals the system if a spike in the frequency of related tweets occurs. Then, the stream of social media is filtered based on the textual content. This component also identifies

[1]https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

each sub-event that occurred by type through extracting a list of representative key terms that describe the sub-event.

- Tweets Categorization: Each tweet in the detected sub-event is assigned to a specific team utilizing a hybrid approach. The hybrid approach consists of two phases: a hashtag-based phase and a user-profile phase to categorize the teams' tweets.
- Image Filtering: The stream of social media is filtered based on textual content. In this case, the filter's task is to remove irrelevant and noisy images; moreover, it discards visual data that is not applicable for event summarization.
- Sentiment Classification: Each tweet in the detected sub-event is classified into positive, negative, or neutral categories, utilizing a football-specific sentiment classifier.
- Popularity Prediction Model: The set of images posted during the time of the detected sub-event are passed through the image popularity prediction model in order to predict their popularity ranking score.
- Summary Generator: Produces a chronological multimedia summarization based on the occurrence of sub-events during the event time by amassing a representative set of text and images.

3.2.1 Sub-event Discovery

The majority of previous works rely on detecting sub-events using a fixed, predefined threshold that is calculated based on historical data. In these instances, the predefined threshold depends on the dataset used and may not be effective on another dataset which may differ in the frequency range. This approach is not suitable for real-time sub-event detection due to the lack of historical data. Therefore, we choose to adapt a dynamic-threshold method that is able to adjust the threshold based on tweets frequency during the playing time of a football match. This method is proposed by [55], and it is based on calculating the mean and the standard deviation of the stream volume over a moving time window. This component categorizes the sub-events by extracting the most representative key words describing the type of the sub-event. Fans continue discussing a previously detected sub-event long after its occurrence [138, 94], which may lead to the detection of the same sub-event several times. As such, we employ a merging step to determine if two spikes are discussing the same sub-event based on the vocabulary similarity measure. The full description of this component is illustrated in Chapter 6.

3.2.2 Tweet Categorization

In order to analyze fan reactions, we need to associate tweets to a specific team based on the tweet writers' supported sides. Utilizing only hashtags or key words from within the tweets to assign them to a specific team is not a sufficient solution [132]. We propose a more thorough hybrid approach to assign tweets to a specific team which consists of two phases: a hashtag-based phase, and a user-profile phase. We explain the details of the tweet categorization component in Chapter 6.

3.2.3 Image Filtering

People post all types of images on social media during an event using event hashtags, often without consideration of their images' true relevance to the event [1]. A large portion of these images includes screenshots, adverts, maps, memes, and diagrams which are not applicable for summarization purposes [92, 121]. Previous works that considered multimedia for events summarization precluded memes and reaction images as their goal was to generate an objective summary that simply describes what happened during the events. In this thesis, our goal is to produce subjective multimedia summarizations that encapsulate public reactions and opinions during specific events. Hence, memes and reaction images are considered suitable to our subjective summarization in representing fans' reactions over the course of a given event. In order to filter out unsuitable images, we employ a rule-based filter to discard small and advertisement images. Images of specific types that do not communicate valuable information about the event or fans' feelings and reactions such as screenshots, diagrams, and maps are also filtered out before summarizing the event. Here, we develop an image-type classifier to distinguish images belonging to various types of synthetic and natural images classes. A comprehensive description of this component is detailed in Chapter 6.

3.2.4 Image Popularity Prediction

Previous studies on multimedia-based summarization of social events select a set of representative images to illustrate the event. The selection criteria of representative images are centered on coverage, diversity, and popularity or significance. These criteria can be measured given the availability of a detailed list of all images, including their social signals which indicate the significance of the images, such as the number of retweets, likes, or favorites. Since we are addressing the problem of summarizing an event in a real-time

manner, we lack the social media factors that indicate the popularity of an image with respect to the event, after it happen. Therefore, we employ a popularity prediction model to identify which images will be more popular than others during a new event in order to select a subset of these images for the event representation. The analysis of image popularity phenomena and the development of image popularity prediction model is described in Chapter 4.

3.2.5 Sentiment Analysis

Football fans sentiment is manifested and expressed through their shared tweets over the course of a soccer event. These tweets sequentially reflect the changes and evolution in the fans' state-of-mind depending on the sub-events occurring during the game, e.g, goal scoring, penalties, etc. Aggregating the sentiment conveyed through these tweets help to draw a picture of the feelings that fans are experiencing during a specific football event. In order to identify the sentiment expressed in the fans' tweets, we build a sentiment classifier that is designed specifically for football events. Due to the lack of an appropriate manually labeled dataset, we construct a new sentiment dataset expressly for the football domain, and generate a domain-specific sentiment lexicon using our proposed sentiment dataset. Then, we develop the sentiment classification model utilizing the football-specific dataset. The details of the dataset construction and the development process of the sentiment classifier are introduced and discussed thoroughly in Chapter 5.

3.2.6 Summary Generation

The final component in the proposed framework is employed to select the most representative textual and visual data in summarizing the event. The top popular images based on the predicted popularity score will be selected for the visual data used to illustrate the event. The top ranked tweets based on the sum of the terms' weight will be selected as the textual data used to summarize the event. These steps combined will generate the textual-visual subjective summarization for the event. The details of the selection criteria are discussed along with other components in Chapter 6.

3.3 Summary

In this chapter, we describe the subjective multimedia summarization framework by providing an overview of each component's main task in monitoring and producing an event summary. The next chapters discuss the details of the main components, specifically the image popularity prediction model and the sentiment analysis model development process. In chapter 6, we elaborate on the remaining components of the proposed framework, in detail. We will cover the process from the raw input data to finally generating the subjective multimedia summarization.

Chapter 4

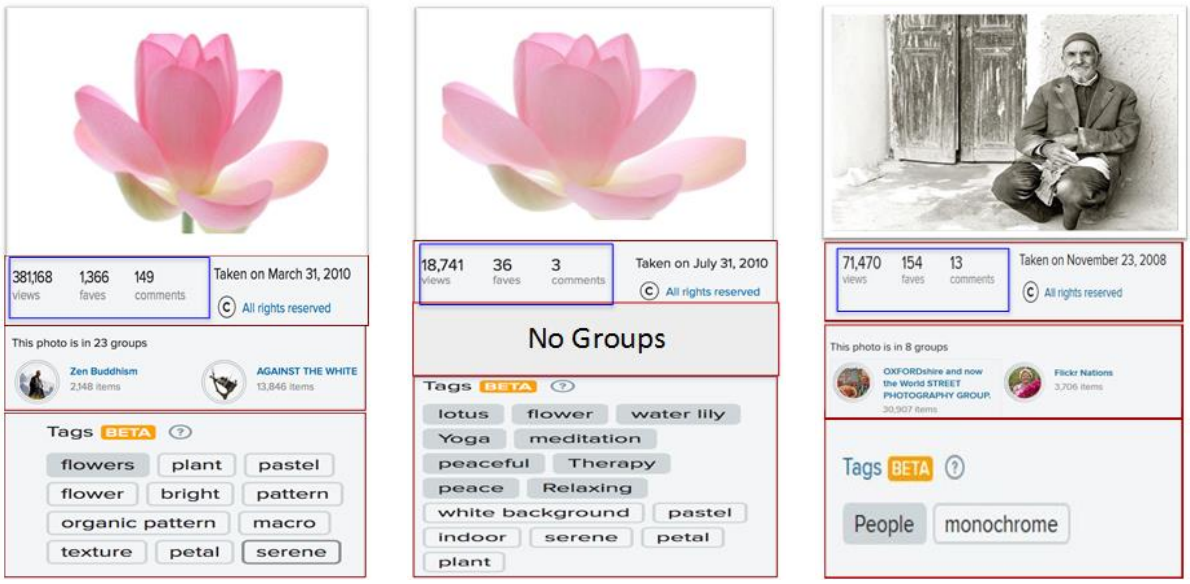
Social Image Popularity Prediction

Over the course of an event, users generate large quantities of images to commemorate the event. Some images garner far more attention than others. Predicting which images will attract most of the spotlight in the early stages of sporting events is imperative for real-time multimedia event summarization. Therefore, we select a sub-set of popular images as representative visual content for the ongoing event. In this chapter, we first discuss the factors that influence social image popularity, followed by a description of our proposed image popularity prediction model.

4.1 Introduction

Image popularity is defined as the level of interaction the image obtains on social media. The interaction can be the number of views, likes, comments, or retweets, depending on the type of social interaction supported by the various platforms. Despite the vastness of social media, we notice that, in general, a limited number of users and their content attract a large share of attention from the online population, while many others go unnoticed. When we consider images in social sharing sites, we witness a considerable variation in their levels of social popularity, regardless of their visual appeal. This variation motivates us to answer the question: which factors impact the popularity of social content?

In this chapter, we explore the factors that may impact the social interactions of users on social media and determine content popularity. We examine Flickr, a photo-sharing social network, in order to observe the social interactions between users and images. Since Flickr contains large banks of publicly available photos, and provides a comprehensive API, it allows us to collect information about images and users along with their social context.



(**a**) (**b**) (**c**)

Figure 4.1: Variation in the number of social interactions between users and images. Visual content, social context, and textual information are important factors for making an image popular

Moreover, Flickr has additional features that allow users to make their photographs accessible and visible to a large number of online users. In addition to sharing images with friends and group members, users on Flickr are encouraged to annotate their images with text-free tags so that they are accessible via keyword searches. As depicted in Figure 4.1, (a), (b), and (c) are examples of images uploaded by the same user. In (a) and (b), the images share the same visual content; whereas in (c), the image represents a different visual concept. The three images in Figure 4.1 received varying amounts of social interactions, even though the images had a similar visual appearance. Figure 4.1 illustrates the differences between the images in terms of surrounding text represented by tags and the number of groups that an image joins, both of which affect the number of social interactions. Attaching descriptive tags to the images and sharing them with suitable groups help in making them popular. This shows how the popularity levels the images are impacted by factors other than visual content. We address the problem by first scrutinizing the social behavior on Flickr to determine how users browse social photo-sharing sites, and then analyzing the correlation between the users and an image's social factors with regard to social popularity. In addition, we investigate the role of associated text in regard to an image's popularity. Next, we propose a system to predict image popularity by implementing a prediction algorithm based on multi-modality features. In the prediction system, in addition to the

social context of users and images, we employ content-based features, and textual features in order to predict the popularity ranking scores of a selected set of images. Our proposed system could be adapted to other platforms considering the differences in the social and contextual features.

4.2 Data Collection and Analysis

We collect images from Flickr to analyze the factors that influence the social interaction between users and images since there is no publicly available dataset with social context information. Data collection procedures are described in Section 4.2.1, and data analysis is discussed in Section 4.2.2.

4.2.1 Data Collection

To collect images, we utilized groups from Flickr, where images are organized by theme and are uploaded by various registered users. One reason to choose groups as a medium to collect images is the differences in image quality and social clues. We selected 31 topics covering a wide range of visual concepts that are listed in the NUS-WIDE dataset [32] and in Flickr's popular tags [1]. Examples of these topics are animal, bike, bridge, cloud, food, football, lake, plane, reflection, tree, wedding, and winter. We used the selected topics as text queries to return lists of groups wherein these visual concepts are represented. For every query, we filtered the list of groups returned to ensure that the groups are public and reflect the concepts. Then, we selected 10 groups for each visual concept, for a total of 310 groups. Images belonging to these groups are downloaded, as is the social information related to these photos. The images and their social context are collected through the Flickr API. Our dataset consists of 1.5 million images uploaded by 90,532 users and we refer to this set as the original dataset. The dataset consists of images with significant variation in the number of views. The maximum number of views an image has received in the dataset is 1,603,158, while 121 images have received the minimum value of views which is equal to zero. The median number of views and the mean are 385 and 1147.96, respectively. Only 0.02% of the images have received views over 100 K, while 16% have obtained less than or equal to 100 views in our dataset. In addition, the mean number of comments and favorite are 12.7 and 16.2 respectively. Similar to the views, the minimum number of comments and favorite an image has received in our dataset is zero. The maximum

[1] https://www.flickr.com/photos/tags

number of comments and favorite an image obtained were 10,541 and 14,694, respectively. 77% of the images in our dataset received less than 10 comments and, following the same pattern, 72% of the images had number of favorites less than 10. Only 0.07% of the total number of images has more than 1000 comments. Similarly, 0.06 % of the images obtain more than 1000 favorite.

4.2.2 Social Interaction Analysis

Social interactions in online social media are classified into explicit or implicit interactions. Explicit interactions require time and effort from the registered users and are denoted by the number of comments and favorites on the Flickr platform. Implicit interactions are represented by the number of views; however, this social metric is not necessarily a reflection of Flickr members' interest. Hence, the gap between explicit and implicit interactions is significant, yet they have a strong positive relationship. This is a valid observation in our dataset, and the correlation between the number of views and the number of comments equals 0.48. From our dataset, we also observe that images which include similar instances or objects may not acquire a similar level of social interactions. In Figure 4.2, we illustrate three examples of visually similar images which, despite the similarity, received different numbers of views: (A) near duplicate images; (B) images share the same instances; and (C) images belong to the same category. This highlights the fact that visibility on social media is not solely dependent on visual content. The following questions arise: If visual content is not sufficient to increase social interaction, what are other factors that influence interactions on social media? How can we leverage these factors that increase image visibility and subsequently boost social interactions? Thus, understanding online social behavior is a fundamental step in the prediction process. In order to address these questions, we analyze the interactions between users and images, considering how users browse the social media sites. This leads to three main points that need to be investigated: Do users find images only by browsing the images uploaded by their contacts or do they find interesting images through survey groups that cluster photographs based on themes? Finally, how does searching for images using keywords affect the visibility of images and the number of social interactions? To find the answers, we construct two samples of images from our original dataset. The first set represents photos with a high number of views; we refer to this set as the representative dataset. The other set of photos contains randomly selected images with a varying number of views to represent the distribution of image popularity; this set is denoted as the random dataset. The descriptions of the two datasets are provided below:

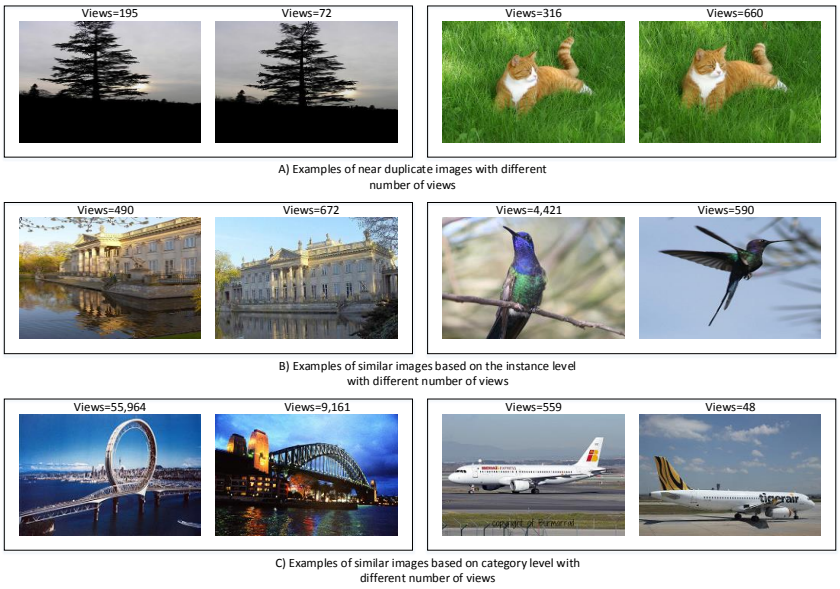


Figure 4.2: Examples of visually similar images. Despite the similarity, they receive a significant different number of views.

- Representative dataset: Images in our original dataset are ranked based on the number of views and then, the Top-1000 photographs are selected to represent a set of the popular images in our dataset.

- Random dataset: We randomly select 50,000 images from our original dataset (1.5 Mio images) that are divided into 10 sets where each set consists of 5000 images. The selected images have varying numbers of views. The most popular image among the 10 sets of images has 1,243,643 views, whereas some images have zero views. In Table 4.1, we provide a brief description of our datasets. For both datasets, we collect the social information of the owners and the images in addition to the metadata of the photos.

Table 4.1: Dataset descriptions.

Dataset	**Description**
Original Dataset	consists of images collected from Flickr groups
Representative Dataset	subset constructed from the original dataset including the ranked Top_1000 based on the number of views
Random Dataset	subset of the original dataset consists of 10 sets each consists of 5000 images randomly selected with different number of views

To begin, because of the limited access to the viewers of images on Flickr, we analyze the source of comments and favorites that images received in both the representative and random datasets. Table 4.2 illustrates the percentage of interactions that images obtained

Table 4.2: Percentage of explicit social interactions with images from groups and user contacts.

Dataset	Interaction	Contacts	Groups
Representative Images	Comment	13.09%	53.1%
	Favorite	17.4%	49.9%
Random Images	Comment	17%	70%
	Favorite	24%	75%

from two sources: groups and user contacts. In both datasets, photos obtained the majority of their interactions from group members. In the representative set, images received 53% of the comments from the groups and only 13% from their owners' contacts. Following the same pattern, the randomly selected images acquired 70% of comments and 75% of favorites from the groups that they joined. We observe that contacts have fewer interactions with images than do group members; however, this can be justified because we only consider contacts that the owners of images are following, as Flickr is a unidirectional network. The information of the reverse contacts, i.e., who follows a user, is not directly available through the Flickr API. In addition, overlap between contacts and group members is typical in social media, yet the overlap is minimal in our case because we consider the users' contacts and not their followers. Nonetheless, a small percentage of interactions are from other users who are not in the groups or the contact lists. The results highlight that the majority of interactions are from groups, which signals the importance of sharing images with groups. Notably, groups vary in popularity and activity level, which affect the number of social interactions; thus, we investigate the popularity power of groups and users on image popularity. Furthermore, we explore the impact of assigning tags and text to images to make the image more accessible and visible in search results. This analysis is provided in Sections 4.2.2.1 and 4.2.2.2.

4.2.2.1 Social Context and Social Interaction

The statistics presented in the last section provide some insight into social interactions; however, we need to understand the impact of groups and users popularity on the social interactions. A user's popularity and activity level can be inferred from several social factors. We consider the number of contacts and the mean count of photo views as indicators of a user's popularity, and the number of joined groups and uploaded images as the activity level. For groups, the numbers of members and shared images represent the group's

activity level. In addition, for groups, through analyzing the data, we discovered that an image can be shared with groups that are not in the owner's list. Some of these groups are private groups and accept images by invitation only, so it is not necessary for the image's owner to join the group. Thus, we divide the groups into image groups where the image is shared but the owner of the image is not a member, and user groups to which the user is subscribed. We compute the strength and the direction of the relationship between both types of social interactions and social factors by calculating Spearman's correlation coefficient [126]. The range of the correlation coefficient is from -1 to $+1$, where a correlation ratio close to 1 or -1 indicates a positive or negative relationship, respectively. Correlation values close to zero show a random relationship. To compute Spearman's correlation, we rank the same set of images based on different criteria: social interactions and social factors that are related to the images or to the owners. For example, we rank the images based on the number of views, number of tags assigned to the images, or number of groups the user has joined where this ranking is done independently. The value of the Spearman's correlation is computed using equation (4.1), where n is the number of images in the set and d_i is the square value of the difference between the rankings of the images.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - n)} \tag{4.1}$$

The results obtained using the random dataset are listed in Table 4.3. The reported results are the average calculated over the 10 sets. The table shows a strong correlation between the popularity of the user and the number of social interactions. The mean number of views of a user's photos is strongly correlated with the number of implicit and explicit interactions (0.58 and 0.70, respectively) of a specific image. Moreover, the size of a user's network is represented by the number of contacts, and has a strong positive relationship with the number of views "implicit interaction" (0.44). The same result also applies to comments and favorites "explicit interaction". In addition to user context, the number of groups in which the image is shared is highly correlated with the number of social interactions. The correlation ranges from 0.52 for comments and favorites to 0.48 for views. The correlation between the number of members in image groups and social interactions is higher than with the number of members of the users' groups. In contrast, there is no relationship between the user's level of activity, or number of uploaded photos, and the number of social interactions.

Table 4.3: Average and variance of correlation between social factors and social interactions on the 10 random datasets.

		Explicit Interaction	**Implicit Interaction**
User context	contacts	$0.479\backslash(\pm 1.2 \times 10^{-4})$	$0.445\backslash(\pm 2.5 \times 10^{-4})$
	uploaded images	$0.039\backslash(\pm 1.7 \times 10^{-4})$	$0.072\backslash(\pm 3.7 \times 10^{-4})$
	number of groups	$0.348\backslash(\pm 5.81 \times 10^{-5})$	$0.349\backslash(\pm 9.83 \times 10^{-5})$
	groups members	$0.269\backslash(\pm 1.67 \times 10^{-5})$	$0.278\backslash(\pm 9.84 \times 10^{-5})$
	mean views	$0.589\backslash(\pm 1.2 \times 10^{-4})$	$0.699\backslash(\pm 4.46 \times 10^{-5})$
Image context	number of groups	$0.608\backslash(\pm 5.38 \times 10^{-5})$	$0.587\backslash(\pm 6.09 \times 10^{-5})$
	groups members	$0.524\backslash(\pm 8.29 \times 10^{-5})$	$0.479\backslash(\pm 2.34 \times 10^{-4})$
	number of tags	$0.233\backslash(\pm 1.8 \times 10^{-4})$	$0.402\backslash(\pm 8.39 \times 10^{-5})$

We further list the correlation results on the representative dataset in Table 4.4. We can see that the user's contacts have a greater impact on the number of comments and favorites than on the number of views. The average number of views for a user's images has a moderate correlation with the number of views of the new image, while a random relationship is found with the number of comments and favorites. Image groups positively influence the number of comments, favorites and views. On the other hand, the number of user's uploaded images has a negative relationship with the number of comments and favorites received. From the results, we can see there are differences between social factors correlation with implicit and explicit interactions in each of the random and representative datasets. The number of images in the representative dataset is smaller than the number of images in the random set which, could impact the correlation measurement. Images in the representative dataset belong to users with large number of followers in general, a detail which influences their images' popularity. Unfortunately, we cannot directly crawl the followers' information from Flickr. In addition, we noticed that the owners of these popular images within our dataset are joining relatively small numbers of groups while their images are being shared with more groups where the user is not a member. In general, it is apparent that a user's popularity and the groups to which he belongs have an influence on the image's popularity. Contacts and groups show a higher correlation with explicit interactions than do the views; however, the popularity of a user's images significantly impacts the number of views. This supports our hypothesis that the scope of a user's network and their choice of groups affect the visibility and popularity of their images in social media.

Table 4.4: Correlation between social factors and social interactions for the representative dataset.

		Explicit Interaction	**Implicit Interaction**
User context	contacts	0.125	0.029
	uploaded images	−0.353	0.0135
	number of groups	−0.19	0.04
	groups members	−0.163	0.075
	mean views	0.083	0.255
Image context	number of groups	0.201	0.112
	groups members	0.14	0.076
	number of tags	0.138	−0.018

4.2.2.2 Textual Context and Social Interaction

Another important feature of Flickr is image annotation, more specifically, tags and text associated with images. This textual information is used to index the images and describe their content to increase accessibility when searching via keywords. We analyzed the correlation between the number of tags associated with an image and the number of views and explicit interactions received by the image. The results are showed in Tables 4.3 and 4.4. In the random dataset, the number of tags has a positive correlation with both types of interactions. The correlation between the number of tags and implicit interaction is 0.40. This value decreased to 0.23 when measuring the strength of the relationship between tags and the number of explicit interaction. This is due to the fact that explicit interaction is motivated usually by social relationship rather than finding an image by search, as is the case in implicit interaction. In the representative dataset, we can see there is a negative relationship between the number of tags associated to the image and implicit interaction. This indicates that some of the popular images are receiving views based on the user popularity; However, this does not negate the fact that assigning tags to images is an important factor to increase images visibility to other users and worth more investigation. Hence, we rank the images in the random dataset based on the number of views received in order to analyze the number of tags assigned to popular and unpopular images within this sample. From Table 4.5, it is obvious that the gap in the number of tags assigned to popular and unpopular photos is significant. The number of tags attached to the top-100 viewed images is 2039. In contrast, the 100 least popular images have only 82 tags. The difference between the number of tags associated with the top-500 ranked images and with

Table 4.5: The gap in the number of tags between popular and unpopular images.

Image Rank	Number of Tags	Image Rank	Number of Tags
Top 100	2039	Least 100	82
Top 200	4132	Least 200	146
Top 500	9932	Least 500	490

the 500 least popular images is 9442.

We went a step further and collected new images from Flickr to analyze the text assigned to these images. We searched for images using the same queries that we used to return the group images (31 different visual concepts) and returned photos based on interesting values calculated by Flickr. Through this step, we obtained two sets of images, interesting images and uninteresting images, based on Flickr's interestingness score. Comparable to our findings using our datasets, the difference in the number of tags is significantly high. Figure 4.3 shows a comparison of the number of tags assigned to the 1000 most popular and most unpopular images belonging to the same visual category. Moreover, we analyzed the text used to describe popular and unpopular images and generated tag-clouds to depict the most frequently used words. Figure 4.4 presents the tags associated with popular and unpopular photos; we can see that popular tags are typically common words such as cloud, car, and tree. The majority of unpopular images lack titles and descriptive text and are associated with fewer tags, if any. In general, these tags are more specific, in that they describe specific attributes regarding the image visual content, such as the car's brand. Furthermore, the text is not comprehensive, and only covers a portion of the visual content and the semantics of the images; unlike interesting images that are rich in textual information. Most of the popular images are uploaded with informative titles and descriptions. In addition, the photos are annotated with tags that describe visual content besides human semantic concepts. These tags are more descriptive and include more searchable keywords, which increases the chance that the images will appear in the search results. The same observation applied to our datasets in terms of tags and description quality.

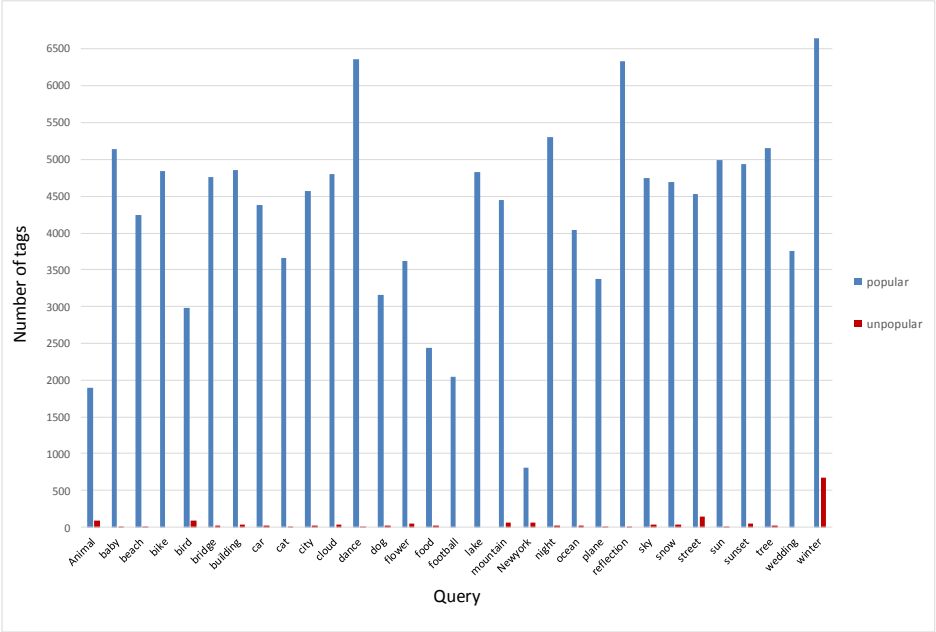


Figure 4.3: The gap between the number of tags associated with popular photos and unpopular photos. The photographs are ranked based on Flickr's interestingness score.

(**a**)text associated with popular photos

(**b**)text associated with unpopular photos

Figure 4.4: Tag-clouds generated from popular and unpopular images on Flickr.

In summary, the data analysis suggests that popularity on social photo-sharing sites is influenced by a user's reachability and the use of the features provided by the service providers. We found out that most of the comments and favorites are from groups members in both datasets. Also, in the random dataset, we can see high correlation with both

types of social interactions and the number of users' contacts. In addition, tags attached to the images are important in finding the images via keywords search. Hence, we can conclude that users browse images and explicitly interact based on their interests and social relationships. At the same time, we cannot neglect that users will use keywords to view images of interest. Here, we conclude that, on Flickr, the number of social interactions an image receives depends on the popularity of the user, the groups selected, and the quality of the text associated with the image. These factors collaboratively increase the popularity of images and play an important role in prediction applications on social media. In the next section, Section 4.3, we demonstrate how we applied our findings with a combination of content features in order to effectively predict the popularity scores of social images.

4.3 Image Popularity Prediction Approach

The popularity of social images is measured by various social signals, depending on the social interactions supported by the social media sites. For example, on Facebook, the popularity can be measured by the number of likes or comments, whereas on Twitter, it is measured by the number of re-tweets. Previous works addressed popularity prediction as a regression or classification problem. In this work, we target Flickr as the main platform for predicting a social image's popularity, where the popularity is related to social interaction behaviors. Consequently, we define an image's popularity based on its received number of explicit or implicit interaction. Explicit interactions are represented by the number of comments and favorites, where users explicitly express their interest in an image. In our work, we refer to these as "interactions". Implicit social interactions are defined simply as the number of views. At this point, we formalize popularity prediction as learning to rank images based on their popularity score in terms of the number of views or the number of comments and favorites. Because our dataset consists of images with dramatic variations in their number of views and interactions, we apply the log function. Images receive social interactions during their time online, so to normalize the effect of the time factor we divide the number of interactions an image has obtained by the number of days since it was first uploaded on Flickr. This is known as a log-normalization approach, which is proposed by [70], and defined in equation (4.2).

$$score_i = log_2 \frac{(p_i + 1)}{T_i} \tag{4.2}$$

where p_i is the popularity measure and T_i is the time duration in days since the uploaded date on Flickr. In the following, we consider two types of popularity measures: (i) "views"

(**a**) cat photos

(**b**) sunset photos

(**c**) a given user photo

Figure 4.5: Variation in the popularity metric "views" for images with similar visual concepts and a user's collection.

is the number of views; (ii) "interaction" is the sum of the number of comments and the number of favorite. Comments and favorite have comparable values and explicitly show the users interests thus we consider them as a measure of popularity.

From our data analysis, we observe that the number of views varies between images within users' collections and groups. Figure 4.5 illustrates an example of this inequality in image popularity scores. Thus, we are proposing a popularity prediction algorithm utilizing multi-modal features. We investigate the effect of different visual features that are designed to represent different visual aspects of images, including visual variances and visual semantics. In addition, we consider the impact of an image's aesthetics, where we hypothesize that if images are similar in terms of visual content and social cues, then the beauty will play an essential role on the popularity of the images. Moreover, we explore the role of contextual and textual factors in predicting an image's popularity. In our approach, we follow the standard framework for prediction, which consists of two main components: feature extraction and model learning. This framework is depicted in Figure

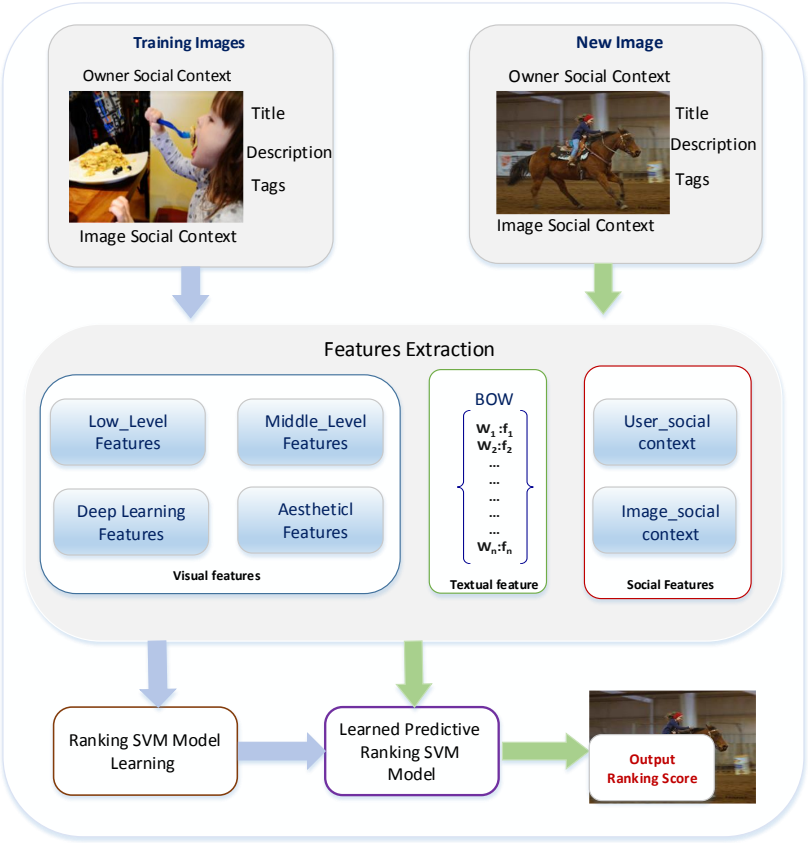


Figure 4.6: The framework for predicting popularity.

4.6. Given a training set of images, we extract different types of features to represent the images. Then, in the model learning stage, we utilize the Ranking Support Vector Machine (Ranking SVM) [66] to be trained on our dataset, and the learned model will be used to predict image popularity ranking score for a new set of photos. In the following sections, we briefly introduce the Ranking SVM algorithm and provide the details of the features that are used in our work.

4.3.1 Ranking SVM

We consider the problem of popularity prediction as a pairwise learning to rank problem. In the pairwise technique, the ranking problem is reduced to a classification problem over a pair of images, where the objective is to learn the preference between the two images. In our experiment, we apply the l_2 regularized l_2 loss function Ranking SVM algorithm to learn the preference between a pair of images with the linear kernel implemented using the LIBLINEAR library [30].

In Ranking SVM, a set of training images with labels is given as $X = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, where x_i is a d-dimensional feature vector of image i and y_i is the popularity score of image i. There exists a preference order for a pair of images such that "x_i is preferred to x_j" is denoted as $x_i > x_j$ when $y_i > y_j$. The objective of a ranking function is to return a score for each data point, an image in our case, where a global ranking over the data is generated. Thus, the ranking function F outputs a ranking score for images such that $F(x_i) > F(x_j)$ for $y_i > y_j$, which minimize the given loss function. F is assumed to be a linear ranking function:

$$F(x) = w \cdot x \tag{4.3}$$

to learn F, the weight vector w should be computed for most of the pairs such that:

$$\begin{aligned} F(x_i) > F(x_j) &\implies (w \cdot x_i > w \cdot x_j) \\ F(x_i) > F(x_j) &\implies w(x_i - x_j) > 0 \end{aligned} \tag{4.4}$$

Now, the relationship between the pair of images x_i and x_j is represented by the new vector $x_i - x_j$. The relationship between any pair of images can therefore be represented by a new feature vector and new labels as follows [53, 24, 66, 51, 154]:

$$x_i - x_j, l = \begin{cases} +1 & y_i > y_j \\ -1 & y_j > y_i \end{cases} \tag{4.5}$$

The problem now becomes a classification problem using SVM that can assign positive or negative label to any vector $x_i - x_j$.

4.3.2 Image Visual Content Features

To investigate the effect of an image's content on its popularity, we consider various types of visual features that describe different perspectives of the image. We use Low level computer vision features that efficiently describe the visual appearance of the image, extracted directly from the pixel information. Although low level features perform well when describing the image, they fail to interpret the semantic of the image. Thus, we leverage middle level features that are designed to address the semantic and affective gaps. In addition to these features, we adopt Aesthetic features to represent the beauty of the image, and we extract high level features that detect the appearance of objects in the image using a deep learning technique.

4.3.2.1 Low level Features

We consider extracting four low level features, described as follows:

- Color Histogram: We extract the RGB color channels to represent the color distribution of the image. The three color histograms are concatenated to form one feature vector of 768 dimensions [122].

- Local Binary Pattern (LBP): A famous texture descriptor widely used in computer vision applications. It works by comparing the value of each pixel with its 8 neighbors in a 3×3 neighborhood. If the value of the selected pixel is greater than its neighbors' values, the neighboring pixels are encoded with 1; otherwise, they take on a value of 0. This results in eight binary numbers that are concatenated in a clockwise direction to calculate the corresponding decimal value of the selected pixel [52, 58]. Ojala et al. [105] recognized that certain patterns are more informative than others, which resulted in the introduction of the uniform LBP. In our work, we used the uniform LBP descriptor, which resulted in 59-dimensional features.

- GIST: A popular descriptor in recognition applications and image retrieval, proposed by [106]. GIST recognizes the scenes in images based on the formation of spatial representations that summarize the description of the image scenes, such as layout and category, with few objects. We adopt the GIST descriptor, where the image is divided into a 4×4 grid, and the orientation of the pixels is computed by a Gabor filter. The use of the GIST algorithm resulted in a 320-dimensional feature vector [106, 37].

- Bag_of_Visual_Word (BoVW): A widely used feature in image classification inspired by the famous Bag-of-Word feature (BoW) used in information retrieval and text mining. Due to the differences between images and discrete words in textual documents, the images are treated as patches of representative samples, where the keypoints are detected by applying the SIFT descriptor. The descriptors are then grouped into clusters, where each cluster represents a quantified visual word. Finally, the images are represented as BoVW vectors based on their vocabulary distribution [103, 20]. In our work, we follow the same approach presented in [17] by considering a 2-layer spatial pyramid and max pooling strategy to generate the BoVW. This resulted in feature vectors of 1500 dimensions.

4.3.2.2 Middle level Features

We have explored middle level features that represent different semantic concepts present in the images. In addition, we consider features that detect the emotions and sentiments that appear in the visual content. The selected features are the following:

- Classemes: A descriptor that is built to detect objects in images. This descriptor is a combination of the outputs of classifiers trained to detect 2659 object categories. These categories consist of visual concepts, including concrete names or abstract nouns, and are suitable for a general-category image search. The category labels are selected from the Large Scale Concept Ontology for Multimedia (LSCOM), which is designed for image retrieval [134].
- Category-Level Attributes: Represents properties shared across categories that can be described using adjectives, such as long and red (in the case of hair), instead of concise object names. In our work, we extracted the attribute features based on the technique introduced in [153], which resulted in a feature of 2000 dimensions.
- SentiBank: Unlike Classemes and Attribute, SentiBank is a middle level feature designed to leverage the affective gap between low level features and the sentiments present in the visual content. SentiBank is designed based on Visual Sentiment Ontology (VSO) and can detect 1200 Adjective-Noun Pairs (ANPs), such as 'peaceful lake' and 'smiling eyes', that are shown in images [17].

4.3.2.3 Aesthetic Features

In addition to semantic concept detection, we investigate the impact of the image's beauty on its popularity. We adopt Aesthetic features that are based on psycho-visual statistics rather than those based on art principles, as proposed by Bhattacharya et al. in [12]. These features are designed for videos, where the features are extracted at the cell, frame, and shot levels. Because we apply the features to images, we only consider cell- and frame-level features. This is a 149-dimensional feature vector.

4.3.2.4 Deep Learning Features

Deep learning techniques have gained greater popularity in image classification and object recognition due to their promising performance. In our work, we used a pre-trained

Convolution Neural Network (CNN) namely the AlexNet model that is trained to classify 1.3 M images from the ImageNet challenge into one of 1000 object categories [75]. More specifically, we used the CAFFE [63] framework of the CNN to extract the following features:

- FC7: We extract features from the last fully connected layer (FC7), which is the layer located directly before the classification layer. The output of this layer is a 4096 dimensional features vector and is considered to be a middle level feature.
- FC8: High level features that are represented by a 1000-dimensional feature vector representing the output of the classification layer, which can distinguish between 1000 objects.

4.3.3 Contextual Features

The data analysis provided in Section 4.2.2 shows a social image's popularity is not only dependent on the visual content, but the contextual factors play a primary role on an image's popularity as well. We define the contextual features as the statistical information about the images and their owners on photo-sharing social networks. Contacts and groups where the images are shared with people interested in similar content show a strong positive influence on the number of views and interaction that the images will receive. Consequently, we must consider different contextual factors that impact the image popularity. We have categorized the contextual features into owners' features and images' features. The owner's contextual factors that are correlated with popularity and considered in our experiments are the number of contacts, total number of uploaded photos, and average number of views of the uploaded images. In addition, we consider the number of groups a user subscribes to, the average number of group members, and the average number of images belonging to the groups. We choose to use the number of group members and the number of images to indicate the popularity and activity level of the groups. Furthermore, we consider the contextual features of images that are provided by the images' owners: the total number of groups, the average number of members participating in these groups, and the average number of photos shared with the groups. The decision to select image groups as a contextual feature was made based on our observations from the data analysis, where we saw that images could be shared with groups that were not in the owner's list. In addition, we include the number of tags associated with the image as a context feature of the image. We decided to exploit the number of tags assigned with images because

intuition led us to believe that an image with more tags will appear more often in search results. We combined all the social features after applying L2-normalization to generate one feature vector with 10 dimensions.

4.3.4 Textual Features

Images are not always given descriptive tags, and the quality of the tags cannot be neglected [69]. This is clearly demonstrated in our data analysis. Thus, we explore the effect of text attached to the image on its popularity. We have used the basic textual feature Bag-of-Word (BOW), which is heavily used in text mining due to its simplicity and good performance. Each image is represented as a feature vector of length n where each element is corresponding to a term in a pre-defined vocabulary. This feature vector can be represented as binary or frequency vector. To generate the vocabulary, we consider two schemes: Term Frequency (TF) to select the most frequent terms appear in images tags, title, and description, and Term Frequency-Inverse Document Frequency (TF-IDF) that reduce the weight assigned to more frequent terms. Before selecting the vocabulary, we applied essential natural language preprocessing steps such as removing the stop words and URLs. Also, we applied the Word-Net Lemmatizer provided in the NLTK [1]. The vocabulary used to generate the BOW is of size 1000 terms. Thus, we have a feature vector with 1000 dimensions. This setting of the vocabulary size is the best choice for our problem based on experimental results.

4.3.5 Fusion Techniques

In previous sections, we presented the features that have been selected to individually predict an image's popularity. In this section, we provide the details of combining multimodal features, where we aim to boost the performance of the prediction algorithm using fusion techniques. We use the following methods:

- Early-fusion: Also known as feature-level fusion. We have normalized the features vectors using L2-normalization and then concatenated the normalized features to generate a single feature vector. Then, the learning stage is performed on the multi-representation vector.

[1] http://www.nltk.org/

- Late-fusion: Integrates the normalized output scores obtained from the learned models based on individual features. In our work, we applied the average late fusion method. This type of fusion is also known as decision-level fusion [125, 7].

- Borda Count method: A rank-based technique that is widely used in meta-search and merging ranked lists. This method is based on a voting process, where each voter ranks a set of n candidates based on his/her preference. For each voter, the top-ranked candidates receive n points, the second ranked candidate receives $n-1$ points, etc. The total number of points for each candidate is calculated from all voters and then used to rank the candidates [5, 112]. Because the outputs of our learned models are based on scores, we transfer them to a rank according to their scores obtained from each model. Then, we can apply the Borda Count method to merge the ranked results.

- Weighted Fusion method: A score-based data fusion approach that is used in information retrieval systems. This method calculates the total score of a document x based on the weight w_i that is assigned to each of the k retrieval system and the score of x obtained by the system[141]. The calculation of the weighted fusion method is given by equation (4.6).

$$S(x) = \sum_{i=1}^{k} w_i \times s_i(x) \tag{4.6}$$

  where the weight w_i is calculated as $w_i = \frac{c_i}{\sum_{i=1}^{k} c_i}$, in which c_i is the Spearman's correlation ratio of system i's performance.

To combine the individual features to improve the performance of the prediction algorithm, we perform the sequential strategy to select the features. In this approach, we rank the features based on the achieved performance correlation. Then, starting from the best feature, we add an additional feature to be integrated based on its rank. If the added feature boosts the performance, we add the next feature until there is no improvement in the performance; when such is the case, we discard the feature that did not generate a further improvement. We select this approach of features combination since we have many different features and it has been used successfully in [65, 64, 135].

4.4 Experiments

In this section, we present the data settings used in our experiments and the results of the discussed algorithms. In addition, we show the effect of combining the multi-modal features. For performance comparison, we compare our proposed approach with different learning methods.

4.4.1 Experimental Setup and Evaluation Criteria

We have conducted experiments in two scenarios: The first scenario depicts the case that occurred in search results where, most often, returned images belong to different users. We denoted the dataset that represents this setting as one_per_user. In the second scenario, we predict the image's popularity within a given user's collection of photos. These scenarios are typical on social media sites and similar to the settings of previous work [70]. The two data sets are sampled from our original dataset, which consists of 1.5 million images. The settings of our datasets are described as follows:

- one-per-user: In this setting, we randomly select one image for each user in our dataset, resulting in a total of 89,663 images. This dataset is divided into 30,000 images used for training our model and the remaining 59,663 images for testing. This setting is suitable for investigating how the differences in visual, social, and textual factors will impact popularity.

- personal-collection: This is a personalized setting. We select the Top-80 users based on the number of images that they contributed to our dataset. The 80 users have in total 155,968 images; for each user, 60% of the photos are used for training, and 40% are used for testing. Collecting all the social context and textual information for all the users and images will grow the data size and due to the limitation in computational resources, we select the top 80 users. We train and evaluate the Ranking SVM algorithm for each user independently. The reported results are the average of the 80 users. In this setting, we follow [70] by discarding the user's contextual factors such as contacts, mean views, and total number of uploaded photos since they are identical for all the images that belong to a specific user. We only consider the contextual features of the images that can be different among images even when they belong to the same user such as: an image groups and tags. This is

because each model is trained for a specific user to predict the popularity of images belong to this specific user.

To evaluate the performance of the prediction algorithm, we use Spearman's coefficient to compute the ranking correlation between the predicted ranking scores of the images and the ground truth scores obtained from Flickr. Spearman's correlation equation was described in Section 4.2.2, equation (4.1).

4.4.2 Effect of Different Features

In this section, we present empirical results that demonstrate the effect of using different modalities on the prediction of an image's number of views and interactions, which, in turn, indicate the social image's popularity. In Section 4.4.2.1, we report the results of using only image content features. A discussion and analysis of the performance of contextual and textual features in the prediction of an image's popularity are provided in Section 4.4.2.2. Furthermore, in Section 4.4.2.3, we investigate the effect of fusing different features on the algorithm's performance, where we observed some improvements.

4.4.2.1 Results of Visual Feature

The results are illustrated in Table 4.6 for the one-per-user and personalized-collection settings. For the one-per-user dataset, low level and Aesthetic features have achieved lower performance than the other features. For low level features, BoVW which is known for its performance in object recognition, has provided the best results in predicting the number of views and interactions. LBP a powerful feature that is used in face detection, achieves a slightly lower rank correlation than BoVW on predicting views and interactions. Color, and Gist features have not help in predicting image popularity. When comparing middle level and deep learning features, both SentiBank and FC7 provide better results than Classemes and Attribute. This highlights the importance of emotional concepts that appear in the image visual content in predicting social image popularity. The difference in the detected visual concepts among middle level features result in the differences in the prediction algorithm performance. Classemes and Attribute detect concrete and abstract visual concepts respectively while SentiBank is designed for detecting sentiment in visual content. We observed that the high level feature FC8 is better at predicting the number of interactions than the number of views. In this setting, Aesthetic feature has achieved

Table 4.6: Performance of visual features on the prediction of popularity for the one-per-user and Personal-collection settings.

		One-Per-User		Personal-Collection	
Features		**Views**	**Interaction**	**Views**	**Interaction**
Low Level	BoVW	0.132	0.142	0.2389	0.1779
	Color	0.093	0.104	0.2299	0.1366
	Gist	0.113	0.093	0.2562	0.1839
	LBP	0.129	0.139	0.357	0.2669
Middle Level	Attribute	0.212	0.212	0.3332	0.2264
	Classemes	0.213	0.207	0.308	0.2206
	SentiBank	0.26	0.264	0.3839	0.2595
Deep Learning	FC7	0.242	0.276	0.4222	0.3025
	FC8	0.215	0.24	0.2142	0.1663
Aesthetic		0.118	0.097	0.3498	0.2569

lower performance than middle and deep learning features. Overall, the middle level and deep learning features outperform low level features.

For the personal-collection setting, LBP has exhibited the best performance in the low level feature category. Specifically, the rank correlation is 0.35 when predicting the number of views, and the correlation is decreased to 0.27 when predicting interactions. Gist feature performance is better than BoVW on this dataset. We have collected the images using text queries thus images belong to a specific user tend to represent similar scene categories. Accordingly, the global Gist descriptor which is designed for scene detection can represent the images more accurately. However, for one-per-user setting, the images have very various visual appearance which affect the Gist performance hence BoVW has obtained better results. For middle level and deep learning features, FC7 performs the best in predicting views and interaction signals, with correlations of 0.42 and 0.30, respectively. Aside from FC7, SentiBank outperforms other middle level and high level features. In the one-per-user dataset, we can see that Aesthetic feature has exhibited a weaker performance than others. This is because the possibility of viewing one image, which is in large pool of images on the Internet, is dictated by many factors, such as the tagged words or the influence of the owner popularity. On the other hand, when comparing the images that belong to one user, beautiful images tend to receive more attention. Thus Aesthetic feature performs better

than others when predicting the popularity of images in the personal-collection dataset. In contrast to the one-per-user setting, it is not only middle level and deep learning features that achieve better results; LBP and Aesthetic features show comparable performances to semantic and object features. FC8 performance is worse than others in the personalized setting. This could be related to that images belong to a specific user usually present similar visual objects. Also, there could be object categories that are not in the FC8 classification, but appear in the images. Moreover, in this setting, implicit interactions are more predictable than explicit interactions.

Despite the differences between the set of images, FC7, SentiBank, and LBP are effectual in predicting image popularity. We can see from the results that there is a variation in the performance of some features, such as Gist, BoVW, and Aesthetic, depending on the set of images. This highlights the importance of using different visual features that represent the images at different levels when predicting the image popularity on Flickr.

4.4.2.2 Results of Contextual and Textual Features

The experimental results of contextual and textual features are shown in Table 4.7. When comparing the results in Tables 4.7 to 4.6, we observe a significant improvement of the prediction algorithm using contextual and textual features over visual features. These results show conclusive evidence that contextual and textual features are effective in predicting images popularity. For the one-per-user setting, the contextual and textual features achieve a better performance when predicting the number of interactions compared to the number of views. Contextual features have a correlation of 0.55, and the textual performance provide a 0.36 rank correlation. When predicting the popularity based on the views, the algorithm performance has decreased to 0.42 and 0.35 using contextual and textual features, respectively. The ability to predict interactions more effectively than views is due to the fact that commenting on images and marking them as favorites are highly dependent on contextual factors such as an owner's contacts and group membership. From the result, we can see that TF and TF-IDF strategies have provided similar results. This is due to that images in this setting are collected from different users and represent various visual topics. Thus, the words used to describe the images are varied depending on the images' visual appearance and the users' preferences.

In the scenario of personal-collection, the performance of textual features, TF and TF-IDF, surpass that of contextual factors in predicting both the number of views and interactions. TF strategy achieves a 0.75 rank correlation in predicting views and 0.52 for

Table 4.7: Performance of contextual and textual features in the prediction of popularity.

		One-Per-User		Personal-Collection	
Features		**Views**	**Interaction**	**Views**	**Interaction**
Contextual features		0.428	0.555	0.3736	0.311
Textual feature	TF	0.353	0.363	0.7589	0.529
	TF-IDF	0.35	0.36	0.446	0.44

interactions. The performance of TF strategy is better than TF-IDF which is opposed the expectation in text classification. This can be explained by that IDF weighting schema assigned lower weight to words that appear more frequent in the text attached to the images which represent the image visual content and semantic. Assigning large weight to rarely appeared terms which do not reflect the images topic, such as "week", "try" and "staying" are not helpful in predicting images' popularity. While more frequent words, such as "tree", "sun" and "swimming" are important for representing the image visual appearance and to predict its popularity. When comparing the performance of contextual features to textual feature (TF strategy), the performance of the algorithm has dropped to 0.37 and 0.31, respectively. In this setting, the decrease in the performance of contextual features is a consequence of utilizing only the image context, more specifically, the image's group information.

4.4.2.3 Results of the Fusion Techniques

In previous sections, we presented the ranking performance results by exploring individual features. In this section, we will provide the results of combining multi-modality features to boost the algorithm's performance. We examine the impact of fusing visual features that represent different aspects of an image's content on the prediction performance, denoted as Visual-Fusion. In addition, we combine visual, contextual, and textual features to evaluate the prediction algorithm's performance. We refer to this as combination as All-Feature-Fusion. The results of the various fusion methods are illustrated in Table 4.8.

For the one-per-user dataset, the fusion of Visual-features results in an improvement of the performance when compared to using individual visual features. The weighted fusion method and late-fusion have achieved slightly better performance than other fusion methods; therein, obtaining a 0.35 rank correlation when predicting the number of views. In predicting the number of interactions, all the fusion methods achieved similar performance

Table 4.8: Performance of fusion techniques in the prediction of images popularity.

		One-Per-User		Personal-Collection	
Fusion Method	**Features**	**Views**	**Interaction**	**Views**	**Interaction**
Early Fusion	Visual-Features	0.3371	0.359	0.5162	0.3868
	All-features	0.5126	0.6292	0.7589	0.5299
Late Fusion	Visual-Features	0.346	0.3566	0.5318	0.3825
	All-feature	0.5192	0.5799	0.7839	0.5683
Borda Count	Visual-Features	0.3371	0.355	0.549	0.407
	All-feature	0.5126	0.569	0.758	0.537
Weighted Fusion	Visual-Features	0.347	0.358	0.533	0.383
	All-feature	0.559	0.601	0.794	0.575

(0.35 rank correlation). When combining visual, contextual, and textual features, the early fusion method exhibits the best performance in predicting the number of interactions obtaining a 0.62 rank correlation. Weighted fusion method has showed a rank correlation of 0.55 when predicting the number of views which is the best among other fusion methods. In the personal-collection, fusing visual features leads to an improvement in the algorithm performance. The Borda count method obtains slightly higher performance than other fusion methods. However, the early fusion of all features as well as the Borda count method each fails to improve the performance over the textual feature performance when predicting the number of views. The best performance is provided by weighted fusion, which obtain a rank correlation of 0.79 and 0.57 when predicting the number of views and interactions, respectively. The results indicate that contextual, textual, and visual features lead to better performance when integrated together. Consequently, the three factors that we have considered in our study complement each other to provide a better prediction performance than simply relying on a single feature model.

4.4.3 Results of Different Learning Methods

In this section, we evaluate the performance of different learning methods by replacing the Ranking SVM with Support Vector Regression (SVR), which is adopted in previous works [70, 43].

The results of SVR model are illustrated in Table 4.9 for the one-per-user and personal-

Table 4.9: Prediction results of SVR model using our features for one-per-user and Personal-collection settings.

		One-Per-User		Personal-Collection	
Features		**Views**	**Interaction**	**Views**	**Interaction**
Low-Level	BoVW	0.029	0.036	0.211	0.143
	Color	0.01	0.018	0.249	0.071
	Gist	0.031	0.024	0.252	0.142
	LBP	0.078	0.1003	0.329	0.228
Middle -Level	Attribute	0.056	0.059	0.308	0.163
	Classemes	0.077	0.076	0.269	0.172
	SentiBank	0.089	0.099	0.381	0.175
Deep Learning	FC7	0.059	0.074	0.349	0.246
	FC8	0.115	0.137	0.207	0.103
Aesthetic		0.064	0.068	0.328	0.231
Contextual features		0.381	0.503	0.323	0.263
Textual feature	TF	0.184	0.215	0.709	0.492
	TF-IDF	0.177	0.202	0.629	0.416

collection settings. In one-per-user setting, while using visual features, the best performance results are achieved by FC8, SentiBank, and LBP features. In personal-collection, in addition to SentiBank and LBP, FC7 and Aesthetic have exhibited better performance levels than other visual features. When comparing different modalities features, contextual and textual features outperform the visual features. In general, the results of SVR performance and the results of Ranking SVM which presented in Sections 4.4.2, demonstrate that some visual features, such as SentiBank, LBP, and FC7, are consistently perform better than other features. In addition, contextual and textual features are very effective in predicting an image's popularity. As showed in Table 4.10, we further examine the results of different fusion techniques using SVR method. Combining different levels of visual features successfully enhance the performance of the prediction algorithm over simply using individual features. When using Borda Count method to combine different levels of visual features in order to predict the number of views in one-per-user setting, the performance increases from achieving only 0.115 rank correlation to 0.167. Moreover, the fusion of multi-modal features is better than depending on single modal approach.

Table 4.10: Performance of fusion techniques in the prediction of images popularity using SVR algorithm.

		One-Per-User		Personal-Collection	
Fusion Method	**Features**	**Views**	**Interaction**	**Views**	**Interaction**
Early Fusion	Visual-Features	0.141	0.187	0.446	0.326
	All-features	0.381	0.503	0.739	0.528
Late Fusion	Visual-Features	0.138	0.228	0.485	0.304
	All-feature	0.381	0.503	0.721	0.517
Borda Count	Visual-Features	0.167	0.192	0.519	0.352
	All-feature	0.381	0.503	0.709	0.492
Weighted Fusion	Visual-Features	0.144	0.164	0.495	0.311
	All-feature	0.384	0.504	0.732	0.529

When comparing the performance of SVR model and Ranking SVM model, listed in Tables 4.6 and 4.7, we observe that Ranking SVM exhibits better performance for both dataset settings. This improvement is due to the selection of the prediction model, and not the selection of the features. Since SVR is built to predict the exact popularity score, there is a loss of some accuracy, as opposed to Ranking SVM, which orders the images based on their popularity scores.

We also compare our proposed approach of utilizing multi-modal features with Ranking SVM against using visual and contextual features with SVR model as used in [70]. Results are reported in Table 4.11. Our proposed system performs much better than the baseline method. The improvement is derived from the utility of heterogeneous social sensory data and more powerful learning method.

To sum up, the experimental results show that the utilization of multi-modality features leads to better performance in predicting an image's popularity on Flickr. While contextual and textual features exhibit better performance than visual features since the popularity on social photo-sharing sites is impacted by other factors than visual content, we cannot neglected the visual content of the image when predicting popularity. Consequently, adopting a multi-modality approach in order to predict image popularity on photo-sharing social networks is more effective than single model approach.

Table 4.11: Comparison of our proposed approach to Baseline method.

	One-Per-User		Personal-Collection	
Features	**Views**	**Interaction**	**Views**	**Interaction**
Proposed approach	0.5192	0.5799	0.7839	0.5683
Baseline [70]	0.09	0.119	0.458	0.188

4.5 Summary

In this chapter, we analyzed the social interaction behavior between online users and social images. Our analysis demonstrated that features provided by social networks facilitate the visibility of social images and boost the social interactions. Sharing images with groups and annotating them with descriptive tags help expand their visibility and reach more online users. In addition, evidence shows that, most popular images have a title and description that describe the content and the semantic behind the photographs; whereas, uninteresting images neglect this aspect. Moreover, a user's popularity influences the popularity of their images, where users with larger social networks and popular images will receive more interactions than inactive users. Notably, a user's images are not all equal in their popularity score. Thus, we propose to predict an image's popularity on Flickr by considering three factors: image content, user and image contextual cues and textual features. We conducted extensive experiments on real dataset to evaluate the effectiveness of each individual features as well as their combination. Furthermore, combining multi-modal features boosts the prediction algorithm's performance. Consequently, our proposed method of utilizing multi-modal approach to predict an image popularity on social media is more effective than single modal approach.

Chapter 5

Sentiment Identification in Football-Specific Tweets

5.1 Overview

When football fans commiserate on social media, these "football talks" may sound like a new language for non-fans, especially when inundated with slang terms. Furthermore, sentiment expressed by football fans is often accompanied by the use of expletives which can make the analysis even more challenging [21]. From a sentiment analysis perspective, using a standard sentiment classifier with football conversations could lead to learning confusion. For instance, "That long bomb was sick!" indicates a positive sentiment in the football domain even though the words "bomb" and "sick" are associated with negative sentiment in general context. Similarly, this tweet: "Barcelona did it!!!! Holy s***!!!! VIVA BARCELONA!!!! #ChampionsLeague" in a general sentiment model would be classified as negative due to the inclusion of expletives, yet it conveys a positive sentiment where the fan is cheering for his team. Thus, the question remains - how do we efficiently analyze sentiment expressed by football fans and track the changes during event time?

Prior studies involving sentiment analysis such as [9], [10] and [49] used machine learning and lexicon-based approaches. These three studies are thoroughly explained in the related work section, chapter 2. Fundamental issues which appear in most of the existing works include training sentiment classifiers on a general dataset and extracting features based on general sentiment lexicons. The quality of the classification performance is dependent on determining an effective set of features. This is especially true for lexical features, since one word or sentence may reflect different sentiments within diverse domains [89].

Additionally, the lack of a sufficient manually labeled football dataset resulted in limited progress in football-specific sentiment analysis on social media. Reviewing the literature indicates that the only publicly available football sentiment dataset is the FIFA World Cup 2014 dataset, which was introduced by [9] and [10]. This dataset, however, is automatically labeled. As a result, the data are prone to a noisy label, where a tweet could be mislabeled and assigned to the incorrect class. The following tweets are examples of positive tweets incorrectly labeled as negative:

- "@FraseForster your a classssssss goal keeper"
- "@Cristiano good team"
- "dancing after the goal here in Miami class"
- "BOOOM! Goal of the tournament by @Tim_Cahill - what a legend. Come on you Aussies!"

Likewise, these tweets are categorized as positive, while they obviously express negative sentiment:

- "F*** NOOOOOO! !!! WTF!!!!! S*** NETHERLANDS BREAKS THE DRAW WITH A 2ND GOAL"
- "No f**** way that's a penalty. Cheaters win again.This is why I stopped watching football man. Use a bloody video ref."
- "Omf the referees are destroying the #WorldCup"
- "Bullshit decision. Referees are f*** up this world cup."

Hence, in this chapter, we tackle the challenge of identifying fan sentiment during a sporting event, specifically a football event. We propose a sentiment classification model that is designed specifically for football posts on social media. We first introduce the process of constructing our football-specific sentiment dataset, which consists of tweets related to popular football events and is annotated manually by human beings. The following sections present the pipeline for building a football-specific sentiment analysis model.

5.2 Football-Specific Sentiment Dataset

This section provides a detailed description of our data collection process and data annotation procedure. The goal of this dataset is to provide a benchmark dataset for the football domain where researchers can utilize this dataset for sentiment analysis and comparison purposes.

5.2.1 Data Collection

We use Twitter as our data source in building our corpus. To ensure that tweets are related to football games, we have collected tweets that were posted during two popular football events: the FIFA World Cup 2014 and the UEFA Champions League 2016/2017.

5.2.1.1 FIFA World Cup (FIFA) 2014

The FIFA World Cup 2014 dataset was collected from Twitter by [9] and [10] during the period between June 6th and July 14th 2014, using the Twitter Streaming API. The tweets were filtered by official hashtags, teams hashtags and user names of teams and players. Each tweet was automatically labeled by polarity using Aylien API[1]. The resulting list of tweets only included the tweets ids and the polarity labels. Thus, we have used the Twitter Search API to retrieve the tweets information. Accordingly, we have obtained 440,917 English tweets. The distribution of tweets over the classes is as follows: 192,717 positive, 133,158 negative, and 115,042 neutral tweets.

5.2.1.2 Champions League (CL) 2016/2017

To collect tweets related to CL 2016/2017, we have considered the official hashtag '#Championsleague' as a seed to retrieve tweets related to this event. The Twitter Search API was used for the period of June 1st, 2016, to June 15th, 2017, which is the duration of the targeted event. As a result, a total of 380,579 tweets in multi-languages were obtained. To enrich our data coverage of the CL event, we have ranked the hashtags that appear in our set based on their occurrence in tweets. Then, the top 15 ranked hashtags have been selected to retrieve more data from Twitter in the same period of time. Consequently, we have collected 2,811,833 tweets, where 37% of tweets are in English, 25% of tweets are in

[1] https://aylien.com/text-api/

Spanish and 12% of tweets are in French. In our work, we only considered English tweets for sentiment analysis. To ensure data quality, we filtered out duplication and discard retweets which are indicated by the presence of the RT symbol. Completing this step left us with 819,848 English tweets.

5.2.2 Annotation Process

We annotated our football tweets using crowdsourcing. This platform for creating a benchmark dataset for different tasks is fast, cheap and scalable [15]. Moreover, the accuracy is close to the agreement among experts as indicated in the previous study [104]. We used the Figure Eight service[2], previously known as CrowdFlower, to crowdsource the annotation of the FIFA 2014 and the CL 2016/2017 tweets. We randomly sampled 25,000 and 31,000 tweets from the CL and the FIFA datasets, respectively. In the following subsections, we describe the annotation task design, quality control, and annotation results.

- **Task Design**

  The task consisted of reading a tweet related to the FIFA World Cup 2014 or the UEFA Champions League 2016/2017, and evaluating the author sentiment expressed in the tweet (positive, negative, or neutral). The annotators (contributors) were provided with the image URL if it was included in a tweet since it can provide more description. Also, we asked the contributors to select their confidence level in their answer on a 5-point scale (not confident to very confident). For reference, we provided the contributors with examples that show tweets with negative, positive, and neutral sentiment. We also required that all the contributors be familiar with football in order to understand the sentiment that appears in each tweet. We posted 56 jobs on Figure Eight and each job consisted of 1,000 tweets to be annotated by four contributors. Each job contained tweets from either the FIFA or the CL events, to ensure consistency.

- **Quality Control**

  Quality control plays a fundamental role in fine tuning the annotation process and providing high-quality annotated data [79]. To ensure the quality of the annotation process, we interspersed manually labeled test tweets, known as gold questions on Figure Eight, within other tweets during the annotation process, and monitored

[2]https://www.figure-eight.com/

Table 5.1: Statistics of manually annotated Football-Specific sentiment dataset.

Dataset	#Positive	#Negative	#Neutral	Total
FIFA 2014	13,871	9,648	6,546	30,065
CL 2016/2017	9,562	8,663	6,236	24,461
CL and FIFA	23,433	18,311	12,782	54,526

the contributors' performance. Each contributor was given an accuracy score that reflected their accuracy on test questions. When a contributor answered a test question incorrectly, he/she was notified and provided with the right answer immediately. Each contributor was expected to maintain an 80% accuracy score during the whole annotation process. If an individual's accuracy fell below the 80%, the contributor was identified as unreliable and was eliminated from the annotation process. Also, all the answers provided by untrusted annotators were discarded. To reduce annotation bias, we only allowed each contributor to annotate a maximum of 10% of the tweets per job. In this manner, the annotation process was not dominated by a small group of contributors. To ensure that annotators have experiences in annotation tasks, we only enlisted users who had achieved level 2 in experience based on Figure Eight standard.

- **Annotation Aggregation**

  We measured the inter-annotation agreement using the average of the pairwise agreement between the annotators. We obtained an agreement of 51% and 55% for the CL 2016/2017 and the FIFA 2014 collections of tweets, respectively. In constructing the final dataset, we assigned tweets to sentiment categories based on the annotators' agreement on the category and, subsequently, discarded noisy tweets. The football-specific dataset consists of 54,526 tweets in total, where 24,461 tweets belong to the CL 2016/2017, and 30,065 tweets are from the FIFA 2014 event. The distribution of the tweets among the three sentiment classes are illustrated in Table 5.1.

5.3 Sentiment Analysis Method

The purpose of sentiment analysis is to identify the sentiment underlying a given text. In general, sentiment analysis can be divided into three levels: document level, sentence level, and fine-grained level. In our work on sentiment analysis, we focus on sentence level, where

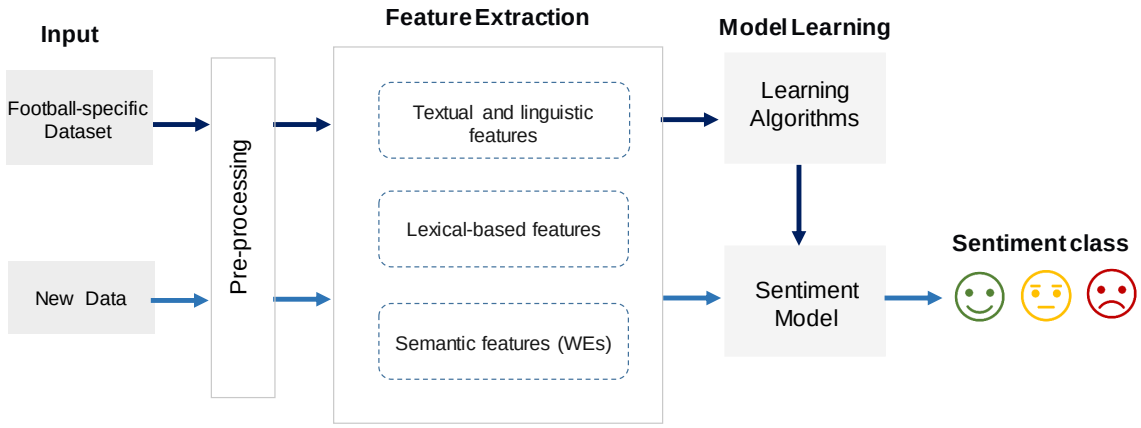


Figure 5.1: General Framework for Sentiment Analysis.

the goal is to determine whether a tweet conveys a positive, negative or neutral sentiment [129]. Sentiment analysis can be considered a document classification problem, aimed at separating documents which express positive and negative sentiments by exploiting certain syntactic and linguistic features [45]. In the literature, the sentiment analysis follows the general framework for the classification problem, which is depicted in Figure 5.1. Feature extraction and classifier learning are the main components of study. Feature extraction is the primary step that converts the raw textual data into representative feature vectors. These features are then fed into a learning algorithm to learn the classification model. In the following sections, we provide the details of the features adopted in our work and briefly introduce the learning algorithms that are employed to learn the sentiment classifier.

5.3.1 Features

We explore the utilization of different types of features, such as the Bag-of-Words feature, which is typically used in sentiment analysis. We also extract lexicon-based features using various existing general sentiment lexicons. Additionally, we develop a new sentiment lexicon oriented for football-specific data, and extract features based on the new lexicon. The following subsections present the features utilized in our work and the process of creating our football-specific sentiment lexicon.

5.3.1.1 Bag-of-Words (BOW)

BOW is one of the most popular representations of textual data and is widely used in text classification. Given a predefined set of vocabulary $V = \{w_1, w_2, ..., w_n\}$, generated using a word or a sequence of words, a document $d \in D$ is represented as an N-dimensional feature vector $X = \{x_1, x_2, ..., x_n\}$. Each element x_i in the feature vector corresponds to a word w_i in the vocabulary. The value of x_i can be a binary value that indicates the appearance of the word w_i in the document d_i or the number of occurrence of w_i in d_i that indicates its term frequency (TF). Term Frequency-Inverse Document Frequency (TF-IDF) is another feature representation that reduces the weight assigned to more frequent words appear in the documents' collection. TF-IDF is a popular and successful representation that shows improvement over TF and is calculated as shown in equation (5.1):

$$\text{TF-IDF}_{(w_i)} = TF_{(w_i)} \times log\frac{|D|}{DF_{(w_i)}} \tag{5.1}$$

where $TF_{(w_i)}$ is the number of occurrence of w_i, $|D|$ is the number of documents in the corpus, and $DF(w_i)$ is the number of document containing the term w_i.

Despite the simplicity and efficiency of BOW, it ignores the co-occurrence of words in texts. To incorporate the word-order, the BOW representation is extended to N-gram language model. In N-gram model, the document is represented as N consecutive words extracted from the collection of documents [87]. The common setting of N-grams is $n \leq 3$. We investigate the impact of using a 2-gram (Bi-gram), 3-gram (Tri-gram), and a combination of different grams: Uni-gram+Bi-gram, Uni-gram+Tri-gram, and Bi-gram+Tri-gram. We use TF-IDF schema for the features vector representation.

5.3.1.2 Word Embedding (WE)

Recently, Word Embeddings (WEs) have proven effective in text classification and sentiment analysis problems. WEs are dense, real-valued, vector representations of words or documents learned from large text corpus using a neural network in an unsupervised manner [23, 155, 157]. Their effectiveness and popularity among researchers is related to their ability to capture the syntactic and the semantic relations between words [113]. Moreover, WEs represent words with lower dimensional vectors compared to the popular BOW representation. Several language models and huge datasets are required to train WEs.

The `Word2Vec` is a shallow, two-layer neural network model proposed by Mikolov et al. [96]. There are two models of Word2Vec to train the embeddings: the *Continuous Bag-*

of-Word (CBOW) model and the *Skip-Gram* model. Given a word context, the CBOW model will predict the word; whereas, the Skip-Gram model predicts the context of a given word. Both models are dependent on a fixed local context window.

A common issue of using a pre-trained WEs is the Out-of-Vocabulary (OOV) problem. This occurs when words in the dataset have no representation in the pre-trained embeddings model [150]. One of the strategies to resolve this issue is to randomly initialize the vector of OOV words [150]. In another approach, Bojanowsk et al. [16], at Facebook, developed a word embedding technique that utilizes a sub-word information in order to overcome the OOV problem. The proposed fastText model represents a word as the sum of the learned representation of its *n-gram* characters. This approach of word representation enables the model to build representation vectors for rare words, slangs, misspelled words and unseen words in the training dataset [6].

In our work, we extract features utilizing both the Word2Vec and fastText word embedding models in order to compare their performance in sentiment analysis. We evaluate the performances of the CBOW and Skip-Gram forms of Word2Vec and fastText. Each tweet in our dataset is represented as a feature vector by computing the average of its words' embedding representation vectors. The embedding values of Out-of-Vocabulary words are randomly initialized.

5.3.1.3 Part-Of-Speech (POS)

POS features are commonly used in sentiment analysis. POS taggers identify each word in a sentence as a noun, verb, adjective, etc. We use The GATE[3] Twitter POS tagger tool to extract the POS features for each word [34]. The GATE POS tagger uses the Penn Treebank tagset, which consists of 36 different tags. We construct the feature vector for each tweet by calculating the frequency of each POS tag.

5.3.1.4 Existing Sentiment Lexicons

Previously developed sentiment lexicons are greatly limited due to insufficient coverage of sentiment words in a single lexicon. Furthermore, many of the existing lexicons do not contain the abbreviations, emoticons, and slang widely used in social media. Thus, in order to achieve better coverage of sentimental words, we use features based on the following sentiment lexicons:

[3]https://gate.ac.uk/

- Bing Liu's Opinion Lexicon (OP) [57]: Manually constructed lexicon from customer reviews about various types of products. It has a list of 2,006 positive and 4,781 negative words.

- AFINN-111 Lexicon (AFINN) [102]: Based on Affective Norms for English Words (ANEW). AFINN is a manually developed lexicon which includes slang and microblog informal words to address the limitation of the ANEW lexicon. It provides sentiment scores for 2,477 words. The sentiment scores range from 1 to 5 for positive words and from -5 to -1 for negative ones.

- NRC Hashtag Sentiment Lexicon (NRC) [97, 71]: Automatically generated lexicon from 775,000 tweets collected using 30 positive and 47 negative hashtagged words. The NRC lexicon has 54,129 unigrams and 316,531 bigrams automatically labeled based on sentiment.

- The Multi-Perspective-Question-Answering Lexicon (MPQA) [147]: It consists of 8,222 words collected from several resources. Each subjective word is manually tagged with its polarity and intensity.

- Emoticons and Slang Lexicon (Emoticons) [72]: It includes emoticons, social media slangs, and abbreviations that are used to express emotion on social media. The total number of entries in this lexicon is 404 and each term is manually labeled as positive or negative.

The lexicons mentioned above are used to extract two features from each individual lexicon: 1) the number of positive tokens and 2) the number of negative tokens.

5.3.2 New Football-Oriented Sentiment Lexicon

Sentiment lexicons are constructed either manually or automatically. Manually generated lexicons leverage the knowledge of experts to label words or terms based on sentiment polarity. This method of constructing sentiment lexicons is costly and has limited coverage [98]. Automatic approaches for generating sentiment lexicons can be divided into corpus-based or thesaurus-based. The thesaurus-based method relies on the assumption that synonyms possess the same polarity [19, 38]. The idea of the thesaurus-based method is to begin with seed words where the sentiment orientation is known and then expand the list by including synonyms and antonyms of the seed words [78]. The corpus-based method uses a domain-specific dataset, instead of the dictionary, to create a sentiment lexicon.

In building our football-specific lexicon, we follow the corpus-based approach, using our collection of football-related tweets. We first preprocess the tweets, removing stop words, URLs, mentions, hashtags, punctuation marks, and digits. Additionally, we remove keywords that are highly correlated with football events such as football clubs' names and their abbreviations, club nicknames, and event official hashtags such as "#WorldCup". In generating our lexicon, we only consider tweets that belong to positive or negative classes, excluding neutral tweets. We then create a vocabulary set that consists of unique tokens appearing in our dataset. The associated sentiment score for each word in the vocabulary set is calculated by adapting the information-theoretic approach proposed in [77, 78], which is based on the well-known information-theoretic measure (TF-IDF), which evaluates the importance of a word in the textual content. The overall score of a word w_i is calculated using equation (5.2):

$$\begin{aligned} Score_{(w_i)} &= (pos(w_i) - neg(w_i)) \times IDF(w_i) \\ where \quad IDF(w_i) &= log\frac{N}{df(w_i)} \end{aligned} \tag{5.2}$$

The overall score of a word w_i is the difference between its positive and negative score multiplied by its inverse document frequency $IDF(w_i)$. N is the total number of tweets in both positive and negative classes, and df is the document frequency (the number of tweets in which w_i appears). Since we have unbalanced classes, we compute $pos(w_i)$ and $neg(w_i)$ based on its frequency relative to positive or negative class as shown in equations (5.3) and (5.4):

$$pos_{(w_i)} = \frac{freq_(w_i, pos)}{N_{(pos)}} \times N \tag{5.3}$$

$$neg_{(w_i)} = \frac{freq(w_i, neg)}{N_{(neg)}} \times N \tag{5.4}$$

Our final lexicon, to which we refer as the Football-specific sentiment lexicon, has entries of 3,479 words: 1,422 of which are labeled as positive and 2,057 negative.

As with the existing sentiment lexicon, we extract features which count the number of positive and negative words in each tweet from the football-specific lexicon.

5.3.3 Learning Algorithms

In this section, we introduce different learning algorithms, including the Support Vector Machine, Naïve Bayes, and Random Forest, all of which are widely used in text classification.

5.3.3.1 Support Vector Machine Classifier (SVM)

Support Vector Machine (SVM) algorithm has shown a robust performance in a wide variety of applications, including text classification [146]. It is a discriminative classifier that learns the boundary between the classes independently from the feature space. The goal of SVM is to find a hyperplane that maximizes the margin between closest instances of the two classes. The closest instances to the margin are called the support vectors.

Given a training dataset consists of N labeled instances $S = \{(x_i, y_i)\}_{i=1}^{N}$ where x_i is the feature vector and $y_i \in \{+1, -1\}$ is the label of instance i. The linear SVM searches for a hyperplane that separates the given data instances which is driven by a linear function (5.5):

$$F(x) = w^T x - b \tag{5.5}$$

$W = \{w_1, w_2, ..., w_n\}^T$ refer to the scalars of each feature dimension and b is the shift of the hyperplane from the original point.

Then, to find the optimal hyperplane, the following optimization problem needs to be solved:

$$\begin{aligned} arg\min_{w,b} \frac{1}{2}\|w\|^2 \\ \text{subject to } y_i(w \cdot x_i + b) \geq 1, \quad i \in [1, n] \end{aligned} \tag{5.6}$$

SVMs were initially designed for binary classification; thus, several extensions have been proposed to handle the multi-class classification problem. The commonly used method is to decompose the classification problem into n binary classification tasks. One-Versus-One (OVO) or One-Versus-All (OVA) are methods in constructing binary classifiers. The OVO method distinguishes one class from another, and the OVA separates one class from all other classes. In this work, we use a linear SVM as a learning algorithm due to its

robust performance in text classification [39]. We adopt the OVA strategy to handle the multi-classification challenge in this thesis as it performs comparably to a more advanced modeling approach despite its simplicity, as stated by Rifkin and Klautauin [114].

5.3.3.2 Multinomial Naïve Bayes classifier (MNB)

Naïve Bayes is a probabilistic classifier that despite its simplicity, performs well in text categorization [108]. Given a set of classes $C = \{c_1, c_2, ..., c_n\}$ and a set of document $D = \{d_1, d_2, d_3, ..., d_m\}$, and $F = \{f_1, f_2, f_3, ..., f_k\}$ is the set of features that represent a document $d \in D$. The probability of document d_i belongs to a class c is computed using Bayes' rules presented in equation (5.7):

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{5.7}$$

The probability of document d is belonging to each class $c \in C$ is calculated individually, then the document d is assigned to class c with the maximum a posteriori (MAP) class as shown in equation (5.8).

$$c_{(map)} = \arg\max_{c \in C} P(c|d) = \arg\max_{c \in C} \frac{P(d|c)P(c)}{P(d)} \tag{5.8}$$

Naïve Bayes classifier is referred to as naïve because it assumes that each feature f_i in a document d is conditionally independent from other features in the given document. The denominator $P(d)$ is constant given the input in equation (5.8) and does not change so we can drop $P(d)$. Thus, we can write equation (5.8) as the following [86, 67]:

$$c_{(map)} = \arg\max_{c \in C} P(c|d) = \arg\max_{c \in C} p(d|c)P(c)) \tag{5.9}$$

$$c_{(map)} = \arg\max_{c \in C} p(d|c)P(c)) = \arg\max_{c \in C} P(f_1, f_2, ..., f_k|c)P(c) \tag{5.10}$$

The final equation for Naïve Bayes classifier is defined as in (5.11):

$$c_{(map)} = \arg\max_{c \in C} P(c) \prod_{f \in F} P(f|c) \tag{5.11}$$

The Multinomial Naïve Bayes (MNB) classifier is a variant of Naïve Bayes classifier which is used with discrete features such as the counts of words in text classification problem. In our work we adapt the Multinomial Naive Bayes classifier for the sentiment analysis problem.

5.3.3.3 Random Forest (RF)

Random Forest (RF) is an ensemble approach devised by Breiman [18] which combines multiple uncorrelated decision trees. Every decision tree in RF is a fully-grown tree constructed using bootstrapped with replacement training samples and random subset of features. Given N instances in the training dataset and F features, to train each decision tree, a random sample with n instances is selected from the original training dataset with replacement to provide a bootstrapped sample. In the process of growing decision trees, a subset of features f from F is selected and each node searches for the best splitting feature within the selected subset of features. The final classification decision of the RF is made based on majority voting from the predictions of all the individual trees in the forest [18, 62]. The randomization in the feature selection reduces the correlation between the trees and improves the classification power. RF displays a strong performance despite noise and overcomes the over-fitting problem that affects a single decision tree [115].

5.3.4 Deep Learning (DL)

The popularity of Deep Neural Networks (DNN) stems from its ability to generate a generic feature vector automatically, which reduces the dependency on hand-crafted features [128]. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are among the most popular deep neural network architectures [143]. CNNs were invented to address computer vision problems and image alike data, while RNNs were designed for processing sequential data [48]. Several previous studies built sentiment classification systems based on CNN models [35, 117], and achieved a decent performance. However, CNNs are not able to capture the context of a sequence of words [128]. Authors in [151] and [133] found that RNNs are more effective and perform better than CNNs in sentence-level sentiment analysis. Accordingly, we build two sentiment classification models based on RNN architecture.

Traditional RNNs suffer from either exploding or vanishing gradient problems which makes the network training more challenging. Several variations of RNN networks have

been developed to overcome this limitation, such as Long-Short-Term-Memory (LSTM), and Gated Recurrent Unit (GRU) [152]. Both LSTM and GRU introduced the gate mechanism over the RNN architecture in order to be able to learn long-term dependencies. We explore the performance of GRU and LSTM based models in identifying the sentiment expressed in a list of given tweets.

Our neural network consists of the following layers:

- **Embedding Layer**: The input to the network is a sequence of words that is projected to a vector of dimension N using pre-trained embedding. In building our models, we opt for using pre-trained word embedding introduced by [11]. The word embedding is generated by training GloVe model on a set of 330 million tweets gathered from December, 2012 to July, 2016 [11]. We use an embedding sized 300-Dimensional vector to represent each tweet. The initialization weights of the embedding layer is based on the pre-trained embedding.

- **LSTM Layer**: After representing the words as vectors, the sequence of the words is inserted into the LSTM layer. The LSTM is designed with a persistent memory cell and gating units in order to capture long-term dependencies. It consists of three gates: input, forget, and output gates, which control the memorizing, updating, and forgetting the state units when appropriate. Two LSTM layers are included in our NN with 150 units in each layer.

- **GRU Layer**: In the second variation of our NN we use a GRU layer instead of LSTM. The GRU is a simpler version of LSTM while maintaining the same effect. It has two gates instead of three: reset and update gates. The update gate acts as a forgetting and an updating gates at the same time, while the reset gate is responsible for deciding which former information should be used for computing the next state. Our network is composed of two GRU layers, and each one consists of 150 hidden units.

- **Softmax Classification Layer**: Fully connected layer with softmax activation function that outputs the probability distribution over the classes.

Regularization is important to avoid model over-fitting during the training phase. We add Gaussian noise of $\sigma = 0.3$ at the embedding layer to reduce over-fitting and add noise to the data. We also apply a Dropout regularization mechanism that randomly drops out neurons along with their connections in the network to make the model generalize better.

We apply dropout of 0.5 at the embedding layer, and 0.5 at the recurrent connections of the LSTM/GRU layer. $L2$ regularization of 0.0001 were added to the loss function. As an optimizer algorithm, we use Adam with a learning rate of 0.001.

5.4 Experiments and Results

Our objective in this study is to build a sentiment classifier that can identify sentiment expressed, specifically, in football tweets. We have utilized our proposed dataset to train different learning algorithms and compare their performance in different experimental settings. In this section, we present the experiment settings, evaluation metrics and the results of utilizing different learning algorithms.

5.4.1 Experimental Settings and Evaluation Criteria

We evaluate the performance of the sentiment classifiers, by conducting experiments for two scenarios: binary and multi-class classifications. In the binary classification setting, we ignore the neutral class and consider only tweets with positive and negative polarity. For both settings, we use the CL 2016/2017, FIFA 2014, and FIFA-CL set, which consist of all the tweets that belong to the FIFA 2014 and the CL 2016/2017. Each dataset is randomly divided into 60% for training and 40% for testing.

For the Deep Learning approach, we have divided the football dataset into 60% for training the network, 20% for validation purposes, and the remaining 20% for testing.

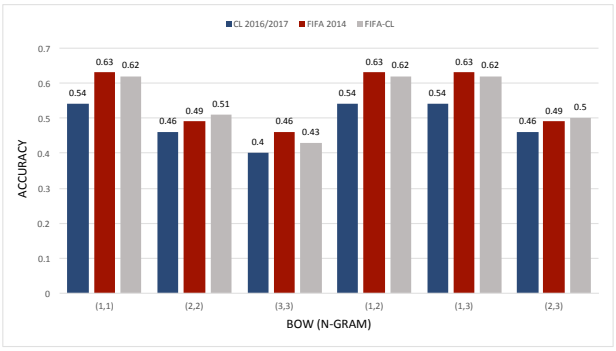
The performance of the sentiment classifiers is measured using accuracy and F-score metrics. Accuracy is defined as shown in equation (5.12) and F-score is calculated as illustrated in equation (5.13).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{5.12}$$

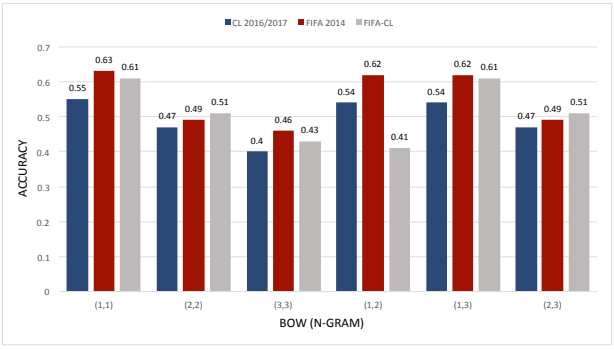
where TP, TN, FP, and FN refer to true positive, true negative, false positive and false negative.

$$F\text{-}score = \frac{2 \times Precision \times Recall}{Precision+Recall} \tag{5.13}$$

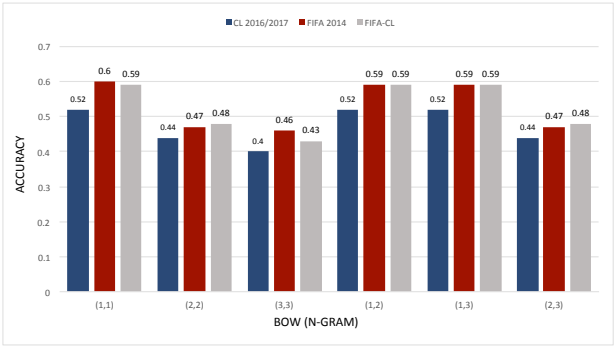
where precision is calculated as $\frac{TP}{TP+FP}$ and recall is defined as $\frac{TP}{TP+FN}$.



(**a**)SVM classifier



(**b**)MNB classifier



(**c**)RF classifier

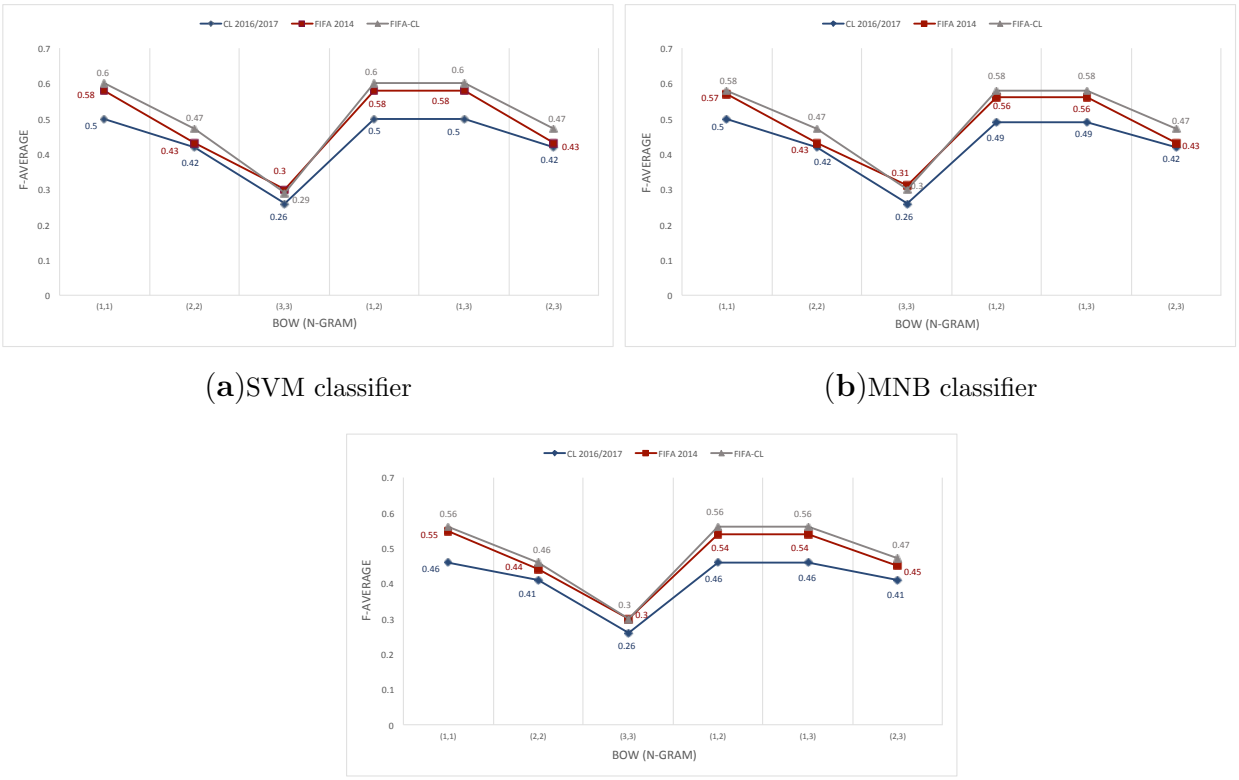
Figure 5.2: Accuracy of different BOW (N-gram) models on CL 2016/2017, FIFA 2014, and FIFA-CL datasets for multi-class classification setting. In (a) the results of SVM classifier, (b) the results of MNB classifier, and in (c) the results of RF classifier.

5.4.2 Experimental Results

In this section, we present empirical results that demonstrate the effect of using different features on the detection of a given tweet sentiment. In section 5.4.2.1, we report the results of using three classes, and in section 5.4.2.2 we discuss the results of using the binary classification setting. Furthermore, in section 5.4.2.3 we investigate the generalization capability of the sentiment model using a cross-dataset setting. Finally, the results of the deep learning-based models are illustrated in section 5.4.2.4.

5.4.2.1 Multi-class classification Results

In this section, we will discuss the performance of different features in categorizing tweets into one of the three sentiment classes: positive, neutral, or negative. First, we investigate



(**a**)SVM classifier

(**b**)MNB classifier

(**c**)RF classifier

Figure 5.3: Average F-score of different BOW (N-gram) models on CL 2016/2017, FIFA 2014, and FIFA-CL datasets for multi-class classification setting. In (a) the results of SVM classifier, (b) the results of MNB classifier, and in (c) the results of RF classifier.

the impact of BOW and N-gram models on the performance of SVM, MNB, and RF sentiment classifiers. The results are illustrated in Figures 5.2 and 5.3 regarding accuracy and F-score, respectively. The experimental results show that Uni-gram achieved better performance than Bi-gram and Tri-gram when used individually. The SVM algorithm obtained an accuracy of 63% when trained on the FIFA 2014 dataset using Uni-gram feature. This accuracy decreased to 49% and 46% using Bi-gram and Tri-gram, respectively. From the results, we can see that the Tri-gram model performance is the least effective with respect to other N-gram models. This is due to the limited number of a tweet's characters and the fact that appearances of Tri-grams are rare in tweets. The combination of Bi-gram+Uni-gram, and Tri-gram+Uni-gram has boosted the performance of Bi-gram and Tri-gram models with respect to accuracy and F-score of SVM, MNB and RF classification models, and among the three datasets. In comparing the performance of SVM, MNB and RF, the results show that SVM has achieved the best performance. The difference in performance accuracy and F-score between SVM and MNB is not significant. This

observation is consistent among CL 2016/2017, FIFA 2014, and FIFA-CL datasets. The N-gram models, in general, perform better on FIFA 2014 and FIFA-CL datasets than on the CL 2016/2017, where they include more training instances. Therefore, having sufficient data for training is important for improving the classification model performance.

Table 5.2 illustrates the results of different lexicons, Word Embeddings (WE), and POS features. We also include the best results from the BOW model for comparison purposes. The results show that WEs features have yielded the best performance for all the classifiers except for MNB on all the three datasets. Both Word2Vec and fastText have provided comparable performance. For instance, the SVM achieved an accuracy of 57% when trained on CL 2016/2017 using Word2Vec (W2V), as opposed to 54% utilizing BOW. Similarly, the performance of RF classifier improved by 2% when trained using

Table 5.2: Performance of different features on: CL 2016/2017, FIFA 2014, and FIFA-CL datasets utilizing different learning algorithms.

		CL 2016/2017		**FIFA 2014**		**CL and FIFA**	
Classifiers	**Features**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**
SVM	BOW	0.54	0.50	0.63	0.58	0.62	0.60
	Opinion Lexicon	0.52	0.44	0.61	0.53	0.56	0.49
	AFINN Lexicon	0.39	0.22	0.46	0.29	0.43	0.26
	MPQA Lexicon	0.48	0.41	0.56	0.49	0.52	0.45
	NRC Lexicon	0.48	0.41	0.55	0.48	0.51	0.45
	Emoticons Lexicon	0.40	0.24	0.47	0.32	0.44	0.28
	Football Lexicon	0.51	0.43	0.55	0.49	0.55	0.48
	POS	0.43	0.37	0.47	0.42	0.46	0.40
	Word2Vec (W2V)	0.57	0.57	0.65	0.61	0.63	0.63
	FastText (FT)	0.56	0.58	0.65	0.61	0.63	0.63
MNB	BOW	0.55	0.50	0.63	0.57	0.61	0.58
	Opinion Lexicon	0.50	0.42	0.51	0.40	0.47	0.35
	AFINN Lexicon	0.39	0.22	0.46	0.29	0.43	0.26
	MPQA Lexicon	0.47	0.39	0.47	0.32	0.44	0.28
	NRC Lexicon	0.40	0.25	0.46	0.29	0.43	0.26
	Emoticons Lexicon	0.40	0.24	0.47	0.32	0.44	0.28
	Football Lexicon	0.51	0.43	0.56	0.48	0.54	0.46
	POS	0.42	0.35	0.46	0.31	0.43	0.28
	Word2Vec (W2V)	0.52	0.52	0.57	0.54	0.56	0.54
	FastText (FT)	0.51	0.51	0.54	0.49	0.54	0.50
RF	BOW	0.52	0.46	0.60	0.55	0.59	0.56
	Opinion Lexicon	0.52	0.44	0.60	0.53	0.56	0.49
	AFINN Lexicon	0.39	0.22	0.46	0.29	0.43	0.26
	MPQA Lexicon	0.48	0.41	0.56	0.49	0.52	0.45
	NRC Lexicon	0.48	0.41	0.54	0.48	0.51	0.45
	Emoticons Lexicon	0.40	0.24	0.47	0.32	0.44	0.28
	Football Lexicon	0.51	0.43	0.55	0.49	0.56	0.48
	POS	0.42	0.40	0.44	0.42	0.45	0.44
	Word2Vec (W2V)	0.56	0.56	0.62	0.56	0.60	0.56
	FastText (FT)	0.55	0.55	0.62	0.56	0.60	0.56

WEs compared to the BOW on the FIFA 2014 dataset. MNB shows different results when using WEs features, as it obtains a lower accuracy than BOW. The degradation of the MNB performance could be a result of the scaling step used with WEs features. Among all the features, the AFINN lexicon achieved the worst performance in terms of accuracy and F-score. Similarly, the Emoticons and NRC lexicons show lower performance levels than Opinion and Football lexicons, since many tweets do not include emoticons or explicit emotional hashtags. The Opinion lexicon outperformed other general sentiment lexicons. When comparing the performance of the Football lexicon to other general lexicons, the results illustrate that the Football lexicon achieves similar performance to the best general sentiment lexicon (Opinion lexicon) used in our experiment. Using the Football lexicon, the MNB classifier has obtained an accuracy of 56% whereas the accuracy dropped to 51% using the Opinion lexicon on the FIFA 2014 dataset. The POS feature outperforms the AFINN lexicon in terms of accuracy and F-score. In some cases, the POS feature shows better performance than the NRC and the Emoticons lexicons. The best results when investigating the performance of different features on the three datasets individually comes from the WEs, then BOW (Uni-gram model). Comparing the performance of the learning algorithms on the CL 2016/2017, the FIFA 2014, and the FIFA-CL datasets shows that the SVM outperforms the MNB and RF classifiers.

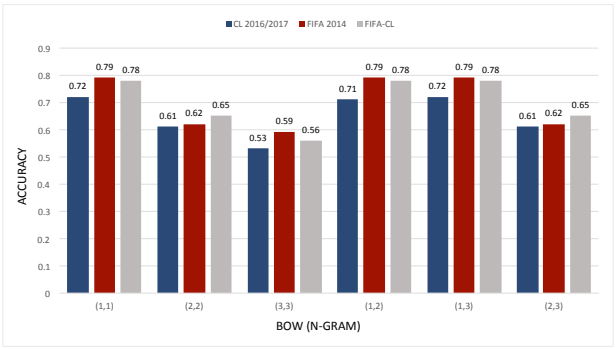
Through the process, we observed that the performance of the different learning algorithms is better when trained on larger datasets. For example, the SVM classifier has achieved an accuracy of 54% on the CL 2016/2017 dataset using BOW. This accuracy increases to 63% when used on the FIFA dataset.

We have presented the sentiment performance results through the exploration of individual features. We will now provide the results of combining different features in order to

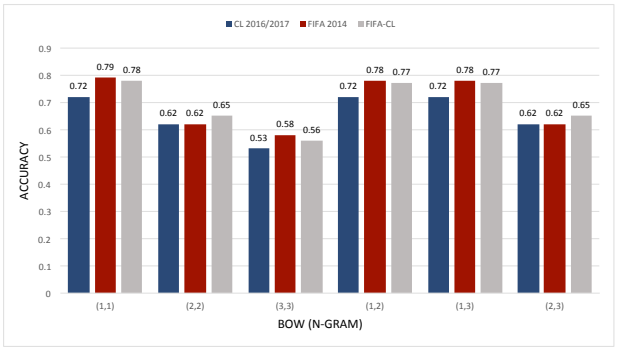
Table 5.3: Performance of features combination on different classifiers using the three datasets.

	SVM						**MNB**						**RF**					
	CL 2016/2017		**FIFA 2014**		**CL and FIFA**		**CL 2016/2017**		**FIFA 2014**		**CL and FIFA**		**CL 2016/2017**		**FIFA 2014**		**CL and FIFA**	
Features	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**
BOW+ General Lexicon	0.56	0.52	0.64	0.59	0.63	0.61	0.56	0.51	0.64	0.58	0.62	0.59	0.55	0.50	0.63	0.58	0.62	0.59
BOW+POS	0.55	0.51	0.63	0.58	0.63	0.61	0.55	0.50	0.62	0.57	0.61	0.58	0.52	0.46	0.58	0.53	0.59	0.56
BOW+FT	0.57	0.57	0.65	0.61	0.64	0.64	0.55	0.55	0.62	0.57	0.61	0.59	0.55	0.55	0.62	0.55	0.59	0.56
BOW+W2V	0.58	0.58	0.65	0.61	0.64	0.64	0.55	0.55	0.62	0.58	0.61	0.60	0.55	0.55	0.62	0.56	0.60	0.56
POS+FT	0.58	0.58	0.653	0.61	0.64	0.64	0.68	0.51	0.51	0.54	0.50	0.55	0.55	0.55	0.62	0.56	0.62	0.59
POS+W2V	0.58	0.58	0.66	0.61	0.63	0.63	0.69	0.52	0.52	0.57	0.54	0.57	0.55	0.55	0.62	0.56	0.60	0.57
POS+General Lexicon	0.52	0.47	0.61	0.55	0.58	0.52	0.51	0.43	0.58	0.51	0.54	0.46	0.52	0.50	0.61	0.57	0.58	0.56
BOW+Football Lexicon	0.55	0.51	0.62	0.57	0.62	0.60	0.55	0.50	0.62	0.56	0.61	0.58	0.53	0.48	0.59	0.54	0.60	0.56
POS+Football Lexicon	0.51	0.44	0.56	0.50	0.56	0.50	0.51	0.44	0.56	0.48	0.55	0.47	0.50	0.47	0.55	0.51	0.55	0.53
General Lexicon	0.52	0.47	0.62	0.54	0.57	0.50	0.50	0.42	0.58	0.51	0.54	0.46	0.51	0.46	0.60	0.53	0.56	0.50
lexicons+FT	0.58	0.58	0.66	0.62	0.64	0.64	0.52	0.52	0.60	0.53	0.57	0.53	0.56	0.56	0.65	0.58	0.62	0.59
lexicons+W2V	0.58	0.58	0.66	0.62	0.64	0.64	0.53	0.53	0.62	0.55	0.58	0.56	0.56	0.56	0.64	0.58	0.62	0.59
General lexicon+Football lexicon	0.54	0.47	0.61	0.54	0.59	0.51	0.53	0.45	0.61	0.53	0.57	0.50	0.50	0.49	0.58	0.55	0.57	0.55
BOW+POS+General Lexicon	0.57	0.53	0.65	0.60	0.64	0.62	0.56	0.51	0.64	0.58	0.62	0.59	0.54	0.49	0.64	0.58	0.61	0.59
BOW+Football Lexicon+POS	0.56	0.52	0.62	0.58	0.63	0.61	0.55	0.50	0.62	0.56	0.61	0.58	0.53	0.47	0.59	0.53	0.60	0.56
BOW+POS+FT	0.57	0.57	0.66	0.61	0.64	0.64	0.56	0.56	0.62	0.57	0.61	0.59	0.55	0.55	0.62	0.56	0.59	0.55
BOW+POS+W2V	0.58	0.58	0.65	0.61	0.64	0.64	0.56	0.56	0.62	0.58	0.61	0.59	0.55	0.55	0.63	0.56	0.60	0.56
BOW+POS+Lexicons+FT	0.57	0.57	0.66	0.62	0.64	0.64	0.56	0.56	0.63	0.58	0.62	0.60	0.56	0.56	0.65	0.58	0.61	0.57
BOW+POS+Lexicons+W2V	0.58	0.58	0.66	0.62	0.64	0.64	0.56	0.56	0.63	0.59	0.62	0.61	0.56	0.56	*0.65*	0.58	0.61	0.57
BOW+General Lexicon+Football Lexicon+POS	0.57	0.53	0.64	0.59	0.64	0.62	0.56	0.51	0.63	0.57	0.62	0.59	0.55	0.50	0.63	0.57	0.62	0.59

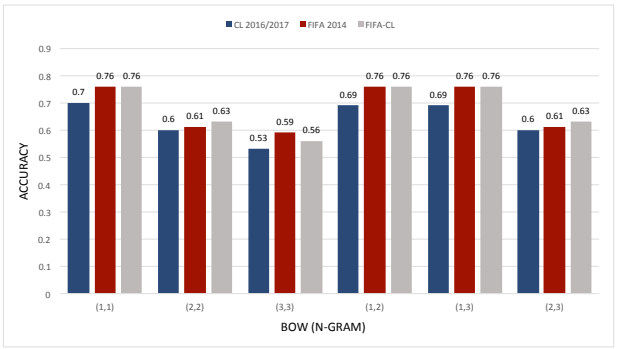
boost the algorithms' performances. We examine the impact of fusing BOW, WEs, lexicons and POS features. The results are illustrated in Table 5.3. Combining features extracted from different general sentiment lexicons boosts the SVM algorithm performance from 61% accuracy, using only the Opinion lexicon, to 62%, on the FIFA 2014 dataset. When fusing general and Football lexicon features, the SVM attained an accuracy of 54% compared to 52% using only the Opinion lexicon on the CL 2016/2017 dataset. The combination of BOW and general lexicons features slightly improved the performance for SVM, MNB, and RF sentiment models. Similarly, combining BOW and Football lexicon features provides a slight improvement in the performance for SVM where the accuracy increased from 54% to $\approx$ 56% on the CL 2016/2017 dataset, see Table 5.3. Moreover, the integration of BOW and WE yields to a slight increase in the SVM performance in terms of accuracy to reach 58% on the CL 2016/2017 dataset. The best performance, in general, is achieved by combining all the features rather than relying on a single type of feature.



(**a**)SVM classifier

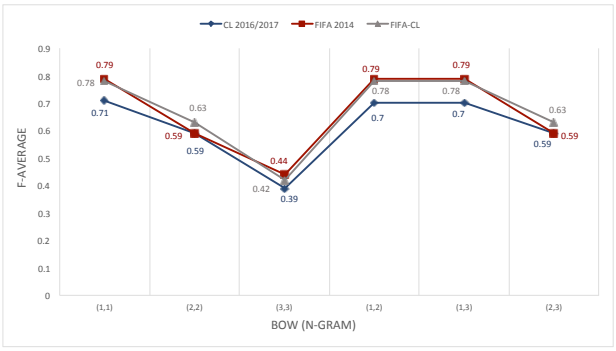


(**b**)MNB classifier

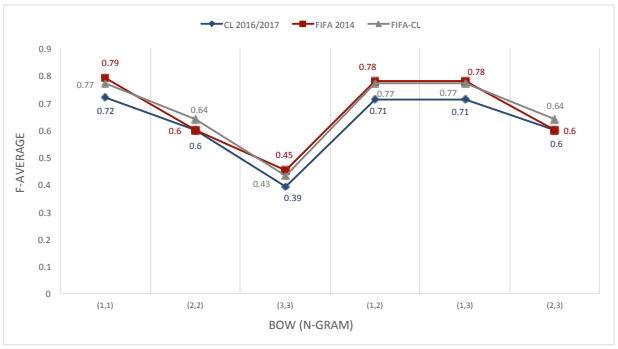


(**c**)RF classifier

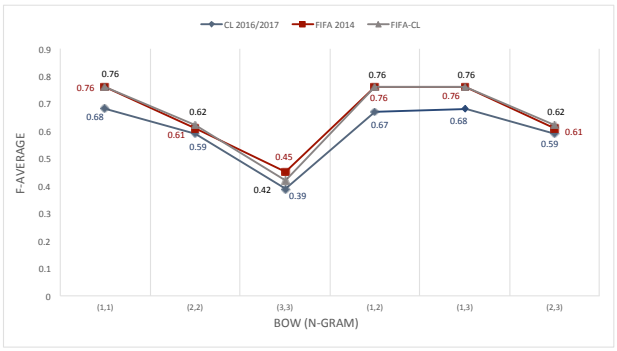
Figure 5.4: Accuracy of different BOW (N-gram) models on CL 2016/2017, FIFA 2014, and FIFA-CL datasets for binary classification setting. In (a) the results of SVM classifier, (b) the results of MNB classifier, and in (c) the results of RF classifier.



(**a**)SVM classifier



(**b**)MNB classifier



(**c**)RF classifier

Figure 5.5: Average F-score of different BOW (N-gram) models on CL 2016/2017, FIFA 2014, and FIFA-CL datasets for binary classification setting. In (a) the results of SVM classifier, (b) the results of MNB classifier, and in (c) the results of RF classifier.

5.4.2.2 Binary Classification Results

We perform experiments on polarity classification using the positive and negative tweets in our constructed datasets. Figures 5.4 and 5.5 report the results of the BOW and N-gram models in accuracy and F-score, respectively. The Uni-gram model has obtained the best performance among the learning algorithms on the three datasets. The Tri-gram model's performance is the worst in comparison with other N-gram models. Bi-gram and Tri-gram models show improvement in their performance when combined with the Uni-gram model. For example, the combination of Uni-gram and Bi-gram results in an accuracy of 79% for the SVM classifier compared to 62% using only the Bi-gram model on the FIFA 2014 dataset. This is similar to the observation in the multi-classification setting where the best performance is obtained using the Uni-gram model. We can see from the results that the learning algorithms perform better in binary classification than in multi-classification task. This is due to the fact that learning algorithms are affected by sentiment class distribution

and perform poorly in rare class.

Comparison of different features' performances in relation to the BOW is shown in Table 5.4. In comparing lexicon features' performances, Opinion and Football lexicons achieved better performances than other lexicons. The MNB algorithm achieved better results using the Football lexicon than the Opinion lexicon. However, the SVM and RF classifiers show better performance using the Opinion lexicon. For instance, the SVM obtains an accuracy of 74% using the Opinion lexicon compared to 72% when using the Football lexicon on the FIFA-CL dataset. The AFINN lexicon performance is the worst among all other lexicons. The limited number of words included in the AFINN lexicon influenced its performance. Interestingly, the POS feature provided better performance than the AFINN lexicon, and

Table 5.4: Performance of different features on: CL 2016/2017, FIFA 2014, and FIFA-CL datasets utilizing different learning algorithms in binary classification setting.

		CL 2016/2017		**FIFA 2014**		**CL and FIFA**	
Classifiers	**Features**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**
SVM	BOW	0.72	0.71	0.79	0.79	0.78	0.78
	Opinion Lexicon	0.69	0.69	0.77	0.77	0.74	0.73
	AFINN Lexicon	0.52	0.36	0.59	0.44	0.56	0.40
	MPQA Lexicon	0.65	0.64	0.72	0.72	0.68	0.68
	NRC Lexicon	0.65	0.65	0.70	0.70	0.67	0.67
	Emoticons Lexicon	0.54	0.40	0.60	0.47	0.57	0.44
	Football Lexicon	0.68	0.68	0.71	0.72	0.72	0.73
	POS	0.58	0.58	0.60	0.59	0.60	0.59
	Word2Vec (W2V)	0.74	0.74	0.82	0.82	0.79	0.79
	FastText (FT)	0.74	0.73	0.82	0.82	0.79	0.79
MNB	BOW	0.72	0.71	0.79	0.79	0.78	0.77
	Opinion Lexicon	0.67	0.65	0.65	0.58	0.62	0.53
	AFINN Lexicon	0.52	0.36	0.59	0.44	0.56	0.4
	MPQA Lexicon	0.63	0.61	0.60	0.48	0.57	0.44
	NRC Lexicon	0.54	0.40	0.59	0.44	0.56	0.40
	Emoticons Lexicon	0.54	0.40	0.60	0.47	0.57	0.44
	Football Lexicon	0.68	0.67	0.71	0.70	0.71	0.69
	POS	0.57	0.55	0.59	0.46	0.57	0.44
	Word2Vec (W2V)	0.69	0.68	0.72	0.71	0.73	0.72
	FastText (FT)	0.68	0.67	0.68	0.68	0.70	0.69
RF	BOW	0.70	0.68	0.76	0.76	0.76	0.76
	Opinion Lexicon	0.69	0.69	0.76	0.77	0.73	0.73
	AFINN Lexicon	0.52	0.36	0.59	0.44	0.56	0.40
	MPQA Lexicon	0.64	0.64	0.72	0.72	0.68	0.68
	NRC Lexicon	0.64	0.64	0.70	0.70	0.67	0.67
	Emoticons Lexicon	0.54	0.40	0.60	0.47	0.57	0.44
	Football Lexicon	0.68	0.68	0.70	0.71	0.73	0.72
	POS	0.56	0.56	0.58	0.58	0.59	0.59
	Word2Vec (W2V)	0.73	0.73	0.80	0.78	0.78	0.77
	FastText (FT)	0.73	0.72	0.79	0.79	0.77	0.77

Table 5.5: Performance of features combination on different classifiers using the three datasets for binary classification setting.

	SVM						MNB						RF					
	CL 2016/2017		FIFA 2014		CL and FIFA		CL 2016/2017		FIFA 2014		CL and FIFA		CL 2016/2017		FIFA 2014		CL and FIFA	
Features	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**	**Accuracy**	**F-Average**
BOW+General Lexicon	0.73	0.73	0.81	0.81	0.79	0.79	0.74	0.73	0.81	0.81	0.79	0.79	0.73	0.73	0.80	0.80	0.78	0.78
BOW+POS	0.72	0.71	0.79	0.79	0.78	0.78	0.73	0.72	0.79	0.79	0.78	0.77	0.69	0.68	0.74	0.74	0.75	0.74
BOW+FT	0.75	0.74	0.82	0.82	0.80	0.80	0.73	0.73	0.79	0.79	0.77	0.77	0.73	0.72	0.79	0.79	0.77	0.77
BOW+W2V	0.75	0.74	0.82	0.82	0.80	0.80	0.73	0.73	0.79	0.79	0.78	0.78	0.73	0.72	0.79	0.79	0.77	0.77
POS+FT	0.74	0.74	0.81	0.82	0.79	0.79	0.68	0.67	0.68	0.68	0.70	0.69	0.73	0.72	0.79	0.80	0.77	0.77
POS+W2V	0.75	0.74	0.81	0.82	0.80	0.80	0.69	0.68	0.71	0.72	0.73	0.72	0.73	0.73	0.79	0.79	0.78	0.77
POS+General Lexicon	0.70	0.70	0.78	0.78	0.75	0.74	0.69	0.68	0.75	0.74	0.70	0.68	0.71	0.71	0.77	0.77	0.75	0.75
BOW+Football Lexicon	0.73	0.72	0.77	0.78	0.78	0.78	0.73	0.72	0.78	0.79	0.78	0.78	0.72	0.71	0.76	0.76	0.77	0.77
POS+Football Lexicon	0.69	0.69	0.72	0.72	0.73	0.73	0.69	0.68	0.71	0.70	0.71	0.71	0.67	0.67	0.71	0.71	0.72	0.72
General Lexicon	0.70	0.69	0.79	0.79	0.75	0.74	0.68	0.66	0.74	0.73	0.70	0.68	0.69	0.69	0.77	0.77	0.74	0.74
lexicons+FT	0.75	0.74	0.82	0.82	0.80	0.80	0.70	0.69	0.77	0.76	0.74	0.73	0.74	0.74	0.82	0.82	0.79	0.78
lexicons+W2V	0.75	0.74	0.82	0.82	0.80	0.80	0.71	0.69	0.78	0.78	0.75	0.75	0.75	0.74	0.82	0.82	0.79	0.79
General lexicon+Football lexicon	0.72	0.72	0.78	0.78	0.77	0.77	0.71	0.70	0.78	0.78	0.75	0.74	0.70	0.70	0.76	0.76	0.76	0.76
BOW+POS+General Lexicon	0.74	0.73	0.81	0.81	0.79	0.79	0.74	0.74	0.81	0.81	0.79	0.78	0.73	0.73	0.80	0.80	0.78	0.78
BOW+Football Lexicon+POS	0.73	0.73	0.78	0.78	0.79	0.79	0.73	0.72	0.78	0.78	0.78	0.78	0.71	0.71	0.75	0.75	0.76	0.76
BOW+POS+FT	0.75	0.74	0.82	0.82	0.80	0.80	0.73	0.73	0.78	0.78	0.77	0.77	0.73	0.72	0.79	0.79	0.77	0.77
BOW+POS+W2V	0.75	0.74	0.81	0.82	0.80	0.80	0.73	0.73	0.78	0.78	0.77	0.77	0.73	0.73	0.79	0.80	0.77	0.77
BOW+POS+Lexicons+FT	0.75	0.74	0.82	0.82	0.80	0.80	0.74	0.74	0.80	0.80	0.78	0.79	0.74	0.74	0.82	0.82	0.78	0.79
BOW+POS+Lexicons+W2V	0.75	0.74	0.82	0.82	0.80	0.80	0.74	0.74	0.80	0.80	0.79	0.78	0.75	0.74	0.82	0.82	0.79	0.78
BOW+General Lexicon+Football Lexicon+POS	0.74	0.74	0.80	0.81	0.80	0.80	0.74	0.74	0.80	0.80	0.79	0.79	0.74	0.74	0.79	0.80	0.79	0.79

in some cases, it also outperformed the Emoticon lexicon. This shows that POS could be helpful in identifying sentiment in a tweet. The best performance among all the individual features is achieved by the WEs (W2V and FT) features. SVM and RF classifiers show an increase in their performance accuracy by 2% to 3% in some cases compared to the best performance achieved using BOW.

Besides exploring the performance of various features individually, we have examined the effect of fusing multiple features and how they perform together on the learning algorithms. The results are illustrated in Table 5.5. Compared to the best single general lexicon feature performance, the fusion of the general lexicons features has improved the performance of the sentiment classifiers by 1% in general. In the case of the MNB algorithm when trained on the FIFA 2014 dataset, the accuracy increases by 9%. The combination of general and Football lexicons features boosted the performance of SVM, MNB, and RF in identifying sentiment in football tweets. The accuracy of the SVM model increases from 69% achieved using only general lexicon to 72% on the CL 2016/2017 dataset. The MNB classifier shows an increase in the performance accuracy by 7% when combining the general and specific lexicons on the FIFA 2014 dataset, and by 4% on the FIFA-CL dataset. The fusion of lexicons and WEs features lead to a slight increase in the performance of SVM by 1% on the CL 2016/2017 and the CL-FIFA datasets. Combining all the features has achieved comparable performance in terms of accuracy to the performance attained by fusing BOW and WEs features for all classifiers on the three datasets.

Table 5.6: performance results of different features on classifying sentiment of FIFA 2014 dataset using models learned from CL 2016/2017 dataset.

	SVM		MNB		RF	
Features	Accuracy	F-average	Accuracy	F-average	Accuracy	F-average
BOW	0.61	0.59	0.60	0.57	0.56	0.53
Opinion Lexicon	0.58	0.51	0.57	0.50	0.58	0.51
AFINN Lexicon	0.46	0.29	0.46	0.29	0.46	0.29
MPQA Lexicon	0.54	0.48	0.50	0.40	0.54	0.48
NRC Lexicon	0.54	0.47	0.46	0.30	0.54	0.47
Emoticons Lexicon	0.47	0.32	0.47	0.32	0.47	0.32
Football Lexicon	0.57	0.52	0.58	0.50	0.58	0.51
Word2Vector (W2V)	0.64	0.63	0.59	0.54	0.61	0.57
FastText (FT)	0.64	0.63	0.58	0.57	0.61	0.56
POS	0.47	0.44	0.48	0.40	0.43	0.43
BOW+W2V	0.63	0.62	0.61	0.58	0.61	0.56
BOW+FT	0.63	0.62	0.61	0.58	0.60	0.55
General Lexicon	0.59	0.53	0.58	0.51	0.58	0.53
BOW+ General Lexicon	0.63	0.61	0.62	0.58	0.61	0.61
W2V+General Lexicon	0.65	0.64	0.62	0.59	0.63	0.61
FT+General Lexicon	0.65	0.64	0.61	0.54	0.63	0.60
BOW+POS	0.61	0.60	0.60	0.57	0.56	0.52
W2V+POS	0.64	0.63	0.59	0.56	0.61	0.57
FT+POS	0.64	0.63	0.58	0.57	0.60	0.56
POS+General Lexicon	0.60	0.56	0.58	0.51	0.58	0.57
BOW+Football Lexicon	0.61	0.59	0.61	0.57	0.59	0.54
POS+Football Lexicon	0.58	0.53	0.58	0.51	0.54	0.53
General lexicon+Football lexicon	0.61	0.55	0.61	0.53	0.57	0.55
BOW+W2V+POS	0.63	0.63	0.60	0.60	0.61	0.57
BOW+FT+POS	0.63	0.62	0.60	0.59	0.60	0.55
BOW+POS+General Lexicon	0.63	0.62	0.62	0.59	0.62	0.60
BOW+Football Lexicon+POS	0.62	0.61	0.61	0.57	0.60	0.56
W2V+POS+General Lexicon	0.65	0.64	0.62	0.60	0.63	0.60
FT+POS+General Lexicon	0.65	0.64	0.61	0.53	0.63	0.60
BOW+General Lexicon+Football Lexicon+POS	0.63	0.63	0.63	0.59	0.63	0.61
BOW+W2V+POS+General Lexicon	0.64	0.63	0.62	0.61	0.63	0.58
BOW+FT+POS+General Lexicon	0.63	0.63	0.62	0.60	0.63	0.58

5.4.2.3 Cross Dataset Performance

Cross-dataset sentiment classification is defined as training a sentiment model on a dataset S_i to predict the sentiment of a tweet t_k in a dataset S_j. We conduct a cross-dataset experiment where we have employed the classification models learned from CL 2016/2017 on FIFA 2014 and vice versa. We have used identical experiment settings as in the previous sections: multi-class and binary classifications.

The results of multi-classification experiments are illustrated in Tables 5.6 and 5.7. Table 5.6 lists the results of different classifiers trained on the CL 2016/2017 dataset and tested on the FIFA 2014. We can see from the results that the accuracy and F-score of the classifiers: SVM, MNB, and RF learned from the CL 2016/2017, Table 5.6, surpass

Table 5.7: performance results of different features on classifying sentiment of CL 2016/2017 dataset using models learned from FIFA 2014 dataset.

	SVM		MNB		RF	
Features	**Accuracy**	**F-average**	**Accuracy**	**F-average**	**Accuracy**	**F-average**
BOW	0.57	0.52	0.57	0.52	0.55	0.51
Opinion Lexicon	0.53	0.46	0.41	0.26	0.53	0.46
AFINN Lexicon	0.39	0.22	0.39	0.22	0.39	0.22
MPQA Lexicon	0.49	0.41	0.39	0.22	0.49	0.41
NRC Lexicon	0.48	0.41	0.39	0.22	0.48	0.41
Emoticons Lexicon	0.40	0.24	0.40	0.24	0.40	0.24
Football Lexicon	0.53	0.45	0.49	0.41	0.53	0.46
Word2Vector (W2V)	0.60	0.57	0.55	0.55	0.57	0.51
FastText (FT)	0.59	0.56	0.52	0.45	0.57	0.50
POS	0.43	0.35	0.39	0.23	0.41	0.39
BOW+W2V	0.59	0.56	0.57	0.53	0.57	0.49
BOW+FT	0.59	0.56	0.57	0.51	0.57	0.49
General Lexicon	0.54	0.46	0.46	0.36	0.54	0.46
BOW+General Lexicon	0.58	0.54	0.58	0.53	0.57	0.53
W2V+General Lexicon	0.60	0.57	0.56	0.52	0.58	0.53
FT+General Lexicon	0.60	0.57	0.54	0.48	0.58	0.53
BOW+POS	0.57	0.54	0.57	0.52	0.53	0.47
W2V+POS	0.60	0.57	0.55	0.51	0.57	0.50
FT+POS	0.60	0.57	0.52	0.46	0.57	0.50
POS+General Lexicon	0.54	0.48	0.48	0.39	0.54	0.51
BOW+Football Lexicon	0.57	0.52	0.57	0.52	0.56	0.49
POS+Football Lexicon	0.54	0.47	0.50	0.42	0.52	0.48
General lexicon+Football lexicon	0.56	0.48	0.54	0.46	0.54	0.52
BOW+W2V+POS	0.60	0.57	0.57	0.55	0.57	0.50
BOW+FT+POS	0.60	0.57	0.57	0.53	0.57	0.50
BOW+POS+General Lexicon	0.59	0.55	0.58	0.53	0.57	0.52
BOW+POS+Football Lexicon	0.58	0.54	0.57	0.52	0.56	0.49
W2V+POS+General Lexicon	0.60	0.57	0.56	0.53	0.58	0.53
FT+POS+General Lexicon	0.60	0.57	0.55	0.55	0.58	0.53
BOW+General Lexicon+Football Lexicon+POS	0.59	0.55	0.58	0.53	0.58	0.53
BOW+W2V+POS+General Lexicon	0.60	0.57	0.58	0.55	0.58	0.51
BOW+FT+POS+General Lexicon	0.60	0.57	0.58	0.54	0.58	0.51

the performance of the classifiers learned from the FIFA 2014 dataset. This is because the performance of the classifiers is affected by the number of instances in each class. The imbalance between classes is larger in the FIFA 2014 than in the CL 2016/2017 dataset, which impacts the performance of the classifiers in classifying tweets that belong to neutral class.

In comparing different individual features' performances, WEs has showed superior performance in relation to other features. Apart from WEs, BOW performance is the best in comparison to other features for all the classifiers. The Football lexicon and Opinion lexicon have exceeded the performance of other lexicons and POS features. It is worth mentioning that the Opinion lexicon has been manually generated, whilst the Football-Specific lexicon has been automatically constructed from football dataset. Likewise, combining the general lexicon features exhibits similar performance in terms of accuracy to the Football lexicon, as the results show in Tables 5.6 and 5.7. Nevertheless, combining general and

Table 5.8: Performance results on FIFA 2014 dataset using models learned from CL 2016/2017 dataset utilizing different features in binary classification setting.

	SVM		MNB		RF	
Features	Accuracy	F-average	Accuracy	F-average	Accuracy	F-average
BOW	0.76	0.76	0.75	0.75	0.72	0.70
Opinion Lexicon	0.75	0.75	0.74	0.73	0.75	0.75
AFINN Lexicon	0.59	0.44	0.59	0.44	0.59	0.44
MPQA Lexicon	0.70	0.70	0.65	0.58	0.70	0.69
NRC Lexicon	0.69	0.69	0.59	0.45	0.69	0.69
Emoticons Lexicon	0.60	0.47	0.60	0.47	0.60	0.47
Football Lexicon	0.74	0.74	0.74	0.73	0.74	0.74
Word2Vec (W2V)	0.80	0.80	0.75	0.74	0.77	0.77
FastText (FT)	0.79	0.79	0.72	0.72	0.77	0.77
POS	0.61	0.61	0.62	0.58	0.57	0.57
BOW+W2V	0.79	0.79	0.77	0.77	0.77	0.77
BOW+FT	0.79	0.79	0.77	0.77	0.77	0.76
General Lexicon	0.76	0.76	0.75	0.74	0.75	0.75
BOW+General Lexicon	0.79	0.79	0.78	0.78	0.78	0.78
W2V+General Lexicon	0.80	0.80	0.77	0.76	0.80	0.79
FT+General Lexicon	0.80	0.80	0.77	0.77	0.80	0.80
BOW+POS	0.77	0.76	0.76	0.76	0.71	0.68
W2V+POS	0.80	0.80	0.74	0.74	0.78	0.77
FT+POS	0.79	0.79	0.72	0.67	0.77	0.76
POS+General Lexicon	0.76	0.76	0.74	0.74	0.75	0.75
BOW+Football Lexicon	0.78	0.78	0.76	0.76	0.76	0.76
POS+Football Lexicon	0.74	0.74	0.74	0.74	0.72	0.72
General lexicon+Soccer lexicon	0.78	0.78	0.78	0.78	0.76	0.76
BOW+W2V+POS	0.79	0.79	0.76	0.76	0.77	0.77
BOW+FT+POS	0.79	0.79	0.77	0.76	0.77	0.76
BOW+POS+General Lexicon	0.79	0.79	0.78	0.78	0.78	0.78
BOW+Football Lexicon+POS	0.78	0.78	0.77	0.76	0.76	0.76
W2V+POS+General Lexicon	0.80	0.80	0.78	0.77	0.80	0.79
FT+POS+General Lexicon	0.80	0.80	0.76	0.76	0.80	0.79
BOW+General Lexicon+Football Lexicon+POS	0.80	0.79	0.78	0.78	0.80	0.79
BOW+W2V+POS+General Lexicon	0.80	0.80	0.78	0.78	0.80	0.79
BOW+FT+POS+General Lexicon	0.80	0.80	0.78	0.78	0.80	0.79

Football lexicons features has boosted the performance of the learning algorithms. The performance of the SVM classifier in Table 5.6 has demonstrated an accuracy of 61% using general and Football lexicons while it achieved 59% using general lexicons individually.

The results of the Binary classification setting is presented in Tables 5.8 and 5.9. Comparing the results of each feature individually, we can see that the highest accuracy is achieved by the Word2Vec feature for SVM and RF classifiers. Analyzing the results which are achieved by the automatically generated lexicon NRC and Football lexicons demonstrates that the domain-specific lexicon (Football) outperforms the general lexicon with respect to accuracy and F-score measures. In addition, the integration of general and

Table 5.9: Performance results on CL 2016/2017 dataset using models learned from FIFA 2014 dataset utilizing different features in binary classification setting.

	SVM		MNB		RF	
Features	Accuracy	F-average	Accuracy	F-average	Accuracy	F-average
BOW	0.74	0.74	0.75	0.75	0.73	0.73
Opinion Lexicon	0.72	0.72	0.55	0.41	0.72	0.72
AFINN Lexicon	0.52	0.36	0.52	0.36	0.52	0.36
MPQA Lexicon	0.67	0.67	0.53	0.37	0.65	0.64
NRC Lexicon	0.64	0.64	0.52	0.36	0.64	0.64
Emoticons Lexicon	0.54	0.39	0.54	0.39	0.54	0.39
Football Lexicon	0.72	0.72	0.66	0.63	0.71	0.71
Word2Vec (W2V)	0.78	0.78	0.72	0.72	0.76	0.76
FastText (FT)	0.78	0.78	0.69	0.69	0.76	0.76
POS	0.57	0.54	0.53	0.37	0.56	0.55
BOW+W2V	0.77	0.77	0.76	0.76	0.77	0.76
BOW+FT	0.77	0.77	0.76	0.76	0.76	0.76
General Lexicon	0.73	0.72	0.62	0.57	0.72	0.72
BOW+General Lexicon	0.76	0.76	0.76	0.76	0.76	0.76
W2V+General Lexicon	0.78	0.78	0.74	0.68	0.77	0.77
FT+General Lexicon	0.78	0.78	0.72	0.72	0.77	0.77
BOW+POS	0.75	0.75	0.75	0.75	0.71	0.71
W2V+POS	0.78	0.78	0.72	0.72	0.76	0.76
FT+POS	0.77	0.77	0.69	0.69	0.76	0.76
POS+General Lexicon	0.73	0.72	0.65	0.61	0.73	0.72
BOW+Football Lexicon	0.75	0.75	0.76	0.76	0.75	0.75
POS+Football Lexicon	0.72	0.72	0.67	0.66	0.71	0.70
General lexicon+Football lexicon	0.75	0.75	0.72	0.71	0.73	0.73
BOW+W2V+POS	0.78	0.78	0.76	0.76	0.76	0.76
BOW+FT+POS	0.78	0.78	0.76	0.76	0.76	0.76
BOW+POS+General Lexicon	0.77	0.77	0.76	0.76	0.76	0.75
BOW+Soccer Lexicon+POS	0.76	0.76	0.76	0.76	0.74	0.74
W2V+POS+General Lexicon	0.78	0.78	0.74	0.74	0.77	0.77
FT+POS+General Lexicon	0.78	0.78	0.72	0.72	0.77	0.77
BOW+General Lexicon+Football Lexicon+POS	0.77	0.77	0.77	0.77	0.77	0.77
BOW+W2V+POS+General Lexicon	0.78	0.78	0.77	0.77	0.77	0.77
BOW+FT+POS+General Lexicon	0.78	0.78	0.77	0.77	0.77	0.77

Football lexicon features has attained better performance than relying on the combination of general lexicons features or individual lexicon. However, fusing both POS and BOW features with general lexicon has raised the accuracy of the SVM classifier by 1% over the combination with the Football lexicon. Here, it worthy to mention that integration of general lexicons includes manually and automatically generated sentiment lexicons and the performance of this combination is very similar to a single automatically constructed specific lexicon. Overall, the best performance of SVM, MNB, and RF classifiers comes from the combination of all the features as illustrated in Tables 5.8 and 5.9. As mentioned before, the result of different sentiment models (SVM, MNB, and RF) trained on the CL 2016/2017 show a more robust performance than the ones trained on the FIFA 2014 dataset. Although the FIFA 2014 includes more training tweets, there is a gap in the number of positive and negative tweets that impacts the classifiers' performance. Conse-

Table 5.10: Performance results of LSTM and GRU based models using the three datasets for multi-class and binary classification settings.

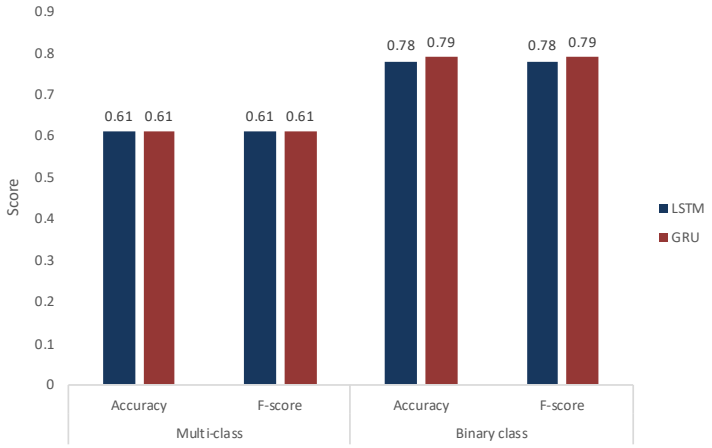
	Multi-class						**Binary class**					
	CL 2016/2017		**FIFA 2014**		**FIFA-CL**		**CL 2016/2017**		**FIFA 2014**		**CL and FIFA**	
Model	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score	Accuracy	F-score
LSTM	0.61	0.58	0.67	0.67	0.65	0.64	0.81	0.81	0.81	0.81	0.81	0.81
GRU	0.60	0.58	0.67	0.66	0.67	0.65	0.80	0.80	0.82	0.82	0.81	0.81

quently, balanced data in the number of instances that belong to each class is important in sentiment analysis.

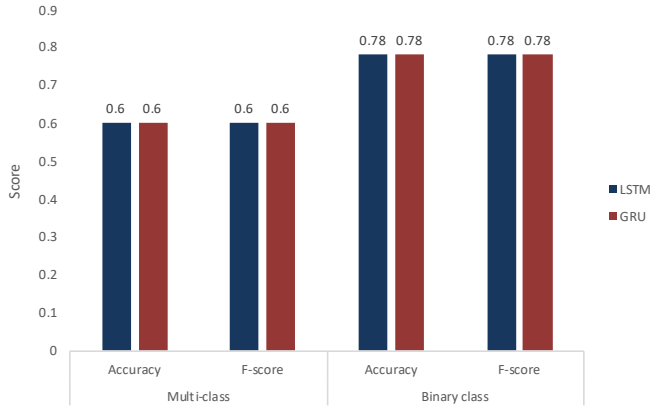
5.4.2.4 Deep Learning Results

In Table 5.10, we report the results of the GRU and LSTM models' performances using the multi-classification and binary settings. The performances of LSTM and GRU are comparable for classifying the sentiment expressed in football tweets. GRU exhibits slightly better performance on the FIFA-CL dataset than LSTM by a 2% increase in the accuracy metric in the multi-class classification. For cross-data setting, the LSTM and GRU models show similar performance in terms of accuracy and F-score as illustrated in Figure 5.6.

When comparing the deep learning models to the traditional machine learning algorithms (SVM, MNB and RF), the deep learning models achieved better accuracy when trained on CL 2016/2017. The SVM classifier obtained an accuracy of 58%, as shown in Table 5.3, while LSTM reached a 61% accuracy on the CL 2016/2017 dataset. For the FIFA 2014 and the FIFA-CL datasets, the GRU model attained slightly better accuracy compared to SVM by 1% and 3% for both datasets, respectively. In the binary classification setting, we can see from the results that the accuracy in identifying sentiment expressed in football tweets significantly increased from only 75% to an accuracy of 81% using LSTM model and 80% using GRU on the CL 2016/2017 dataset. In contrast, the performances of both LSTM and GRU on the FIFA 2014 and the FIFA-CL datasets are similar to traditional machine learning algorithms. In the cross-data setting, the deep learning-based models achieved comparable performance in terms of accuracy and F-score to the SVM performance.



(**a**)Training on CL 2016/2017 and Testing on FIFA 2014



(**b**)Training on FIFA 2014 and Testing on CL 2016/2017

Figure 5.6: Performance results of LSTM and GRU models for cross-data setting. In (a) performance of LSTM and GRU models trained on CL 2016/2017, (b) the results of LSTM and GRU trained on FIFA 2014 and evaluated using CL 2016/2017 dataset.

5.5 Case Study: Sentiment Analysis of Champions League

Football fans' sentiment are manifested and expressed through their shared tweets during a football event. These tweets sequentially reflect the changes and evolution in the fans' sentiment depending on the events of the game, e.g, goal scoring, penalties, etc, as they watch the game. Aggregating the sentiment conveyed through these tweets draw a picture of the feelings that fans are experiencing during a specific football event. Therefore, we leverage our trained sentiment model to analyze fans' sentiment during the UEFA

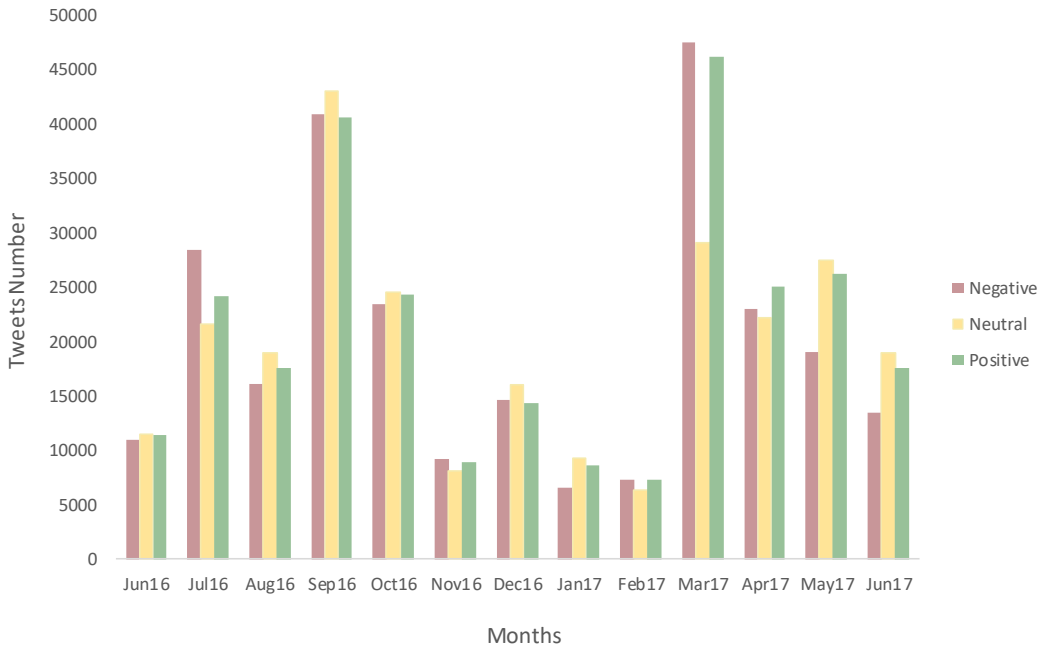


Figure 5.7: Overall sentiment during CL season

Champions League (CL) Soccer 2016/2017 championship. We are particularly interested in tracking the sentiment before, during, and after the games of CL 2016/2017.

5.5.1 Sentiment Analysis During the Season

Figure 5.7 shows the sentiment results of Champions league divided by month. The aggregation of tweet sentiment during a month can reflect the sentiment of people during matches which occurred within this period of time. For example, in March 2017, 8 games were played as part of round of 16 stage. Most of the tweets posted during March 2017 reflect the negative emotion expressed by the teams' fans. Barcelona's defeat of Paris Saint-Germain (PSG) 6-0 on March 8th provoked controversy among fans. PSG fans were elated after their victory by 4 goals on the first leg against Barcelona. They were equally devastated by the subsequent loss since PSG then failed to qualify to Quarter-Finals. On other side, Barcelona fans went from being crestfallen at the possibility of elimination to jubilant when they did, in fact, advance to the Quarter-Finals. In addition, Bayern Munich defeated Arsenal 5-1, which resulted in having more negative tweets than positive ones. Arsenal tweets were shared for days after the game and generated more interactions, as fans were comparing the Arsenal loss to PSG's defeat. The negative impact of Arsenal-Bayern

match results could be related to the higher number of tweets mentioning Arsenal, in that it generated 15,241 tweets, compared to only 5,438 tweets for Bayern. This could be due to the fact we only considered English tweets in our study. Also, the higher number of negative tweets during this month could be explained by the tie result of Atletico-Leverkusen match, which lead to Leverkusen failing to be qualified to Quarter-Finals.

Approximately 70,370 tweets related to the Quarter-Finals were collected. Figure 5.7 shows that the positive sentiment was the dominant during the month of April followed by the negative sentiment. The high level of negative sentiment during this month could be correlated to the heavy loss of Barcelona against Juventus (0-3 for Juventus) on April 11, and the tie game between the same teams on April 19. It is important to mention that Barcelona has the highest number of tweets among all other teams during the Quarter-Finals. The team has 21.57% of the tweets followed by Real Madrid, Juventus, and Leicester (18.22%, 17.48%, and 15.62%, respectively). Hence, the sad feeling of Barcelona fans are obvious when interpreting the results. The negativity of sentiment during the Quarter-Finals could also be attributed to the fact that both Barcelona and Leicester did not qualify for the Semi-Finals. On the other hand, fans of Real Madrid and Juventus expressed mostly positive sentiment, since Real Madrid won both games and Juventus marked a win (3 goals for 0) against Barcelona, and, most importantly, qualified to play in the Semi-Finals.

We have observed that there is a relationship between fan engagements and the social activities of teams. Barcelona and Juventus, for example, use strategies to drive their fans engagements. They publish news and live updates during matches and respond to the fans' posts through their official accounts. Juventus decided to trigger the emotional aspect of their fans by creating the hashtag #ItsTime for the Final stage. This strategy was a success with the fans and the hashtag #ItsTime was used more than their main hashtag #finoallafine (99 times for #finoallafine, 127 times for #ItsTime in our dataset).

5.5.2 Sentiment Analysis of The Final Game

We considered all the tweets related to the final match between Juventus (Juve) and Real Madrid (RM). To extract the tweets, we used the hashtags of the teams and the ones related to the game. Our objective is to examine whether an occurrence of major events during the games would drive fans to engage more than usual on social media. We assume that more social conversations would convey more emotions.

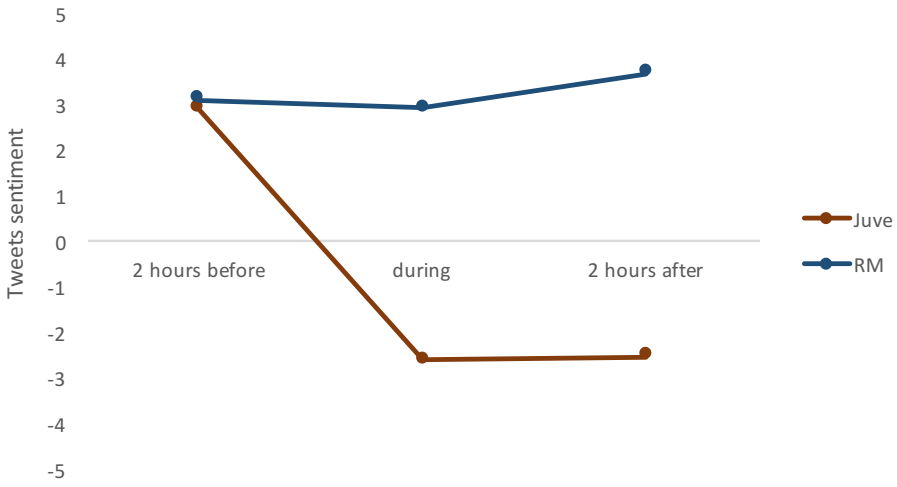


Figure 5.8: Sentiment changes prior, during, and post game in CL Finals

Tracking the sentiment of fan talks throughout the game has revealed interesting insights. Figure 5.8 illustrates the feeling of fans before, during, and after the game. The time window was set to be 2 hours prior to the game starts and 2 hours after game end. Fans of both teams shows positive sentiment leading up to the game kick off. A sample of related tweets shows a lot of support and cheering attitude, hence, the positive sentiment. During the game Juventus fans were not happy about the team performance; whereas supporters of Real Madrid showed a very positive sentiment. This is not surprising since Real Madrid stunned Juventus 4-1 in the Final. When ignoring objective tweets, more than 60% of Juventus tweets contained negative emotion while positive sentiment from Real Madrid's covered 80% of the overall subjective tweets. Post-game sentiment continues the same trend as those logged over the duration of play. Fans of Real Madrid expressed happiness in 74% of the related subjective tweets.

More investigations were conducted on the sentiment during the game. We are particularly interested in examining sentiment changes displayed during occurrences of major events such as scoring goals. We divided the game time tweets into 7 parts each representing a 15-minute portion of the game.

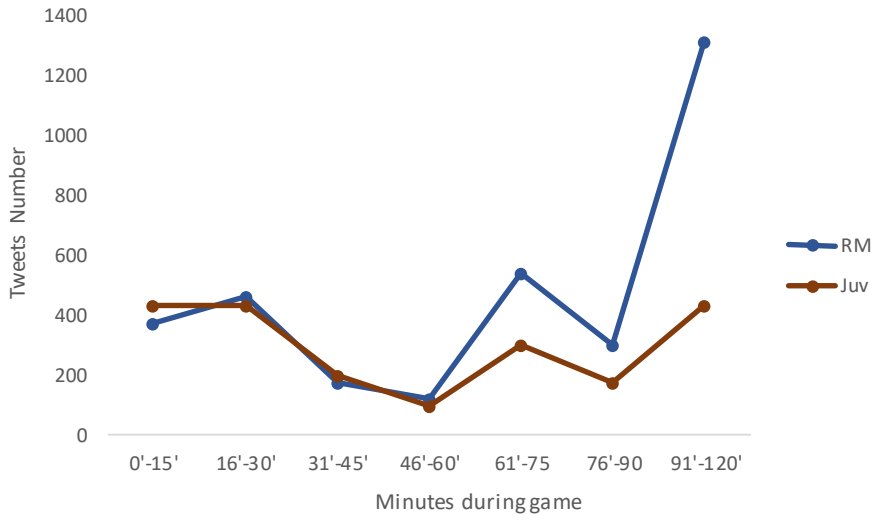


Figure 5.9: Relationship between fans activities and occurrence of goal-scoring events

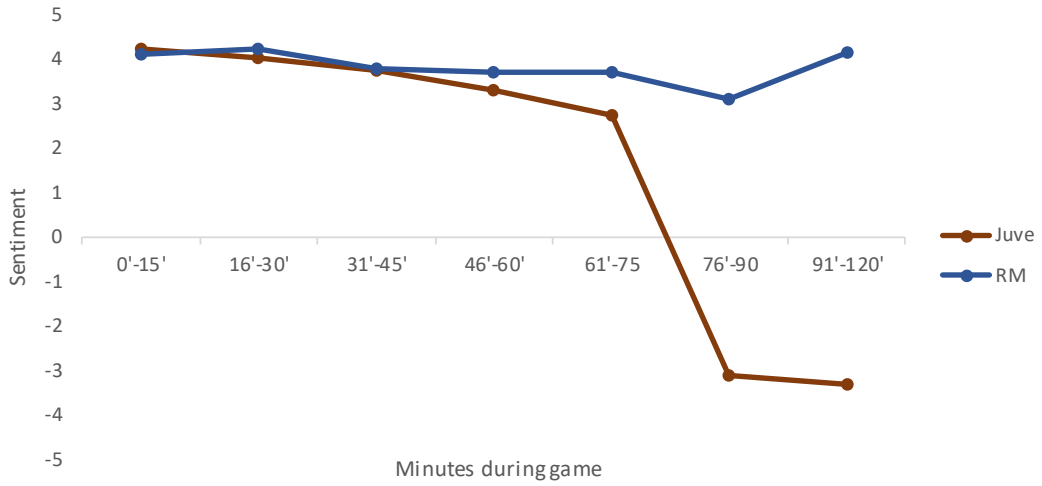


Figure 5.10: Sentiment during the game in CL Finals

We have observed a sudden increase in fans activities when a goal is scored, as shown in Figure 5.9. During the second 15 minutes of the game (16'-30'), Real Madrid scored its first goal. This explains a sudden increase in Real Madrid tweets during the same time

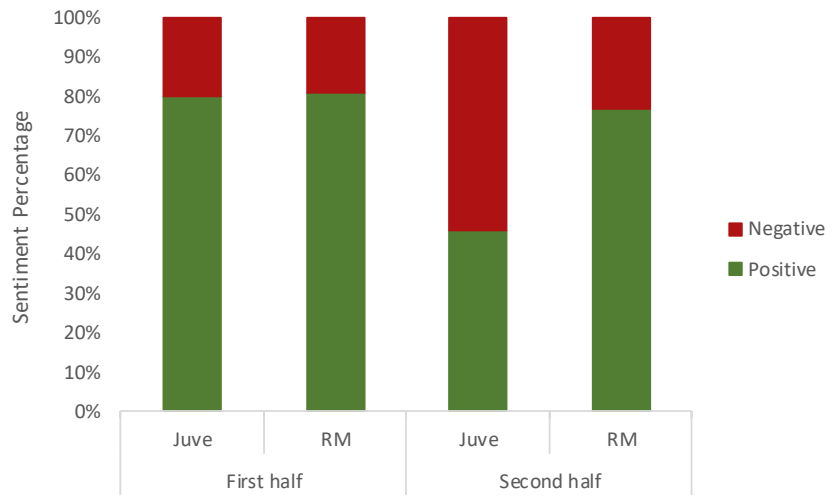


Figure 5.11: Sentiment changes between the first and second halves of the game in CL Finals

point. Both teams were associated to positive sentiment (i.e. 80% convey positiveness) as a result of a 1-1 tie in the first half of the game, as illustrated in Figures 5.10 and 5.11. Real Madrid scored 2 goals between the 60'-65' minute point of the game. Again, this has resulted in another sudden increase in the volume of tweets by fans of both teams. Obviously, Juventus fans expressed negativity against the game's results (see Figure 5.10), including a contrary opinion to the fans of Real Madrid in the second half of the game. The sentiment results in the second half depicted on Figure 5.11, reflect a high level of positiveness in their conversations (more than 75% of their tweets is positive) that dramatically increased (Figures 5.9 and 5.10) after Asensio (RM player) scored the 4th goal on the 90th minute. Note that Fans usually mention both teams in their tweets which, in turn, affect the overall true sentiment representation for each team.

5.6 Summary

In this chapter, we propose a football-specific sentiment dataset which consists of tweets collected from the FIFA World Cup 2014 and the UEFA Champions League 2016/2017 football events. Our dataset consists of 54,526 tweets which are manually labeled by four annotators. We also developed a sentiment lexicon that is oriented for the football do-

main using corpus-based approach. The Football-Specific sentiment lexicon includes a list of 3,479 words labeled according to its polarity. We have conducted an extensive experiment to evaluate the performance of three learning algorithms (SVM, MNB, and RF) and deep learning models in recognizing sentiment appearing in football conversations on social media utilizing different features on our proposed dataset. The results conclude that the Word embeddings (WEs) and BOW model (Uni-gram) achieved the best performance in comparison to other features. In addition, lexicon-based features achieved comparable performances to the BOW. Specifically, the Opinion lexicon and the Football-Specific lexicon gained better performance levels than other lexicons used in our experiments. The SVM, in general, demonstrated robust and consistent performance in comparison with MNB and RF. Moreover, the experimental results have illustrated that training classifiers on datasets with a sufficient and balanced number of instances improves the performance of the learning algorithms.

Chapter 6

Subjective Multimedia Sporting Event Summarization

In chapters 4 and 5, we described the process of developing the popularity prediction and sentiment analysis models, which are the main components for generating subjective sporting summarizations. In this chapter, we will present our methodology during the automatic generation of the sporting event summarizations, specifically, football games.

The digital twin framework for sporting events, illustrated in Chapter 3, Figure 3.1, satisfies the requirements of the digital twins concept proposed in [40, 41]. The digital twin should collect real-time data about the real twin using hard or soft sensors as a data source [40, 41]. In our framework, the social stream obtained from the social media is the data source for the sporting event digital twin, containing valuable data about the event. The digital twins artificial intelligence engine, its main component, extracts knowledge, recognizes patterns, and uncovers latent information from the collected data. Using various artificial intelligence and data mining techniques allow the digital twin to make predictions and provide intelligent suggestions. In our framework, the data obtained from the event will be processed and analyzed using various data mining techniques, including machine learning methods, in order to extract knowledge and hidden information, and understand the real twin circumstances and timeline (a specific sporting event). This is achieved using different components proposed in our framework, including sub-event discovery, tweet categorization, sentiment analysis, and popularity prediction. These components form the artificial intelligence component of the digital twin. The digital twin should be able to communicate with the real world and have representation through various technologies such as virtual reality, hologram, or interaction interface. To allow the sporting events digital twin to interact with the real world, we propose a multimedia representation of the

digital twin through an interactive interface in our framework.

6.1 Problem Formulation

Given a social stream of microblogs $S = \{S_1, S_2, S_3, ..., S_n\}$ related to a football event E_j, where each microblog $S_i = \{P_i, I_i, T_i\}$ consists of textual message P_i, image associated with the text I_i and time of the posted microblog T_i. The P_i and T_i could not be empty, while I_i depends on the existence of the visual part. Our work aims to produce a chronological subjective multi-modal summarization from microblog stream S during the playing time of a given match to update the users with the occurrence of significant sub-events such as goals, red cards, or penalties and how fans react to these sub-events based on their perspectives.

6.2 Sub-event Detection

Many of the previous studies [101, 131] rely on detecting sub-events using a fixed, predefined threshold that is calculated based on historical data. In such a case, the preset threshold depends on the dataset used during the experiments, and may not work on other datasets that vary in the volume of tweets activity. Also, using a fixed threshold is not suitable for real-time sub-event detection due to the absence of the historical data used to determine the threshold in advance. Therefore, we adopt a dynamic-threshold method that can adjust the threshold based on tweets frequency during the playing time of a selected football match. The dynamic-threshold method introduced in [55] is dependent on calculating the mean and the standard deviation of tweets frequency.

To apply the dynamic-threshold method [55], we divide the time sequence $\{t_1, t_2, ..., t_n\}$ of the playing time into a sequence of fixed-size sliding windows $S_1, S_2, ..., S_n$ of length L seconds. Each sliding window has N number of tweets. To determine the dynamic threshold (DTS_i) for S_i at time t_i, we calculate the mean (μ_i) and the standard deviation SD_i of the tweets volume in the previous time sliding windows $\{S_1, S_2, ..., S_{i-1}\}$ as illustrated in equation (6.1). Then, if the tweets frequency at time t_i, $N(S_i)$, is higher than the DTS_i value we consider this time window as a spike which indicates an occurrence of a sub-event. Note that in equation (6.1), the parameter α is used to relax the condition of the DTS and the parameter β is a coefficient. During the experiment, both α and β will be determined.

$$DTS_i = \alpha * (\mu_i + (\beta * SD_i)) \tag{6.1}$$

After detecting the spike in tweet volume, we need to identify the type of sub-event that occurred. To do this, we extract the most representative terms which describe the detected sub-event. In order to compile a list of representative keywords, we first streamline the textual messages by discarding duplicated, irrelevant, and advertisement tweets. We then pre-process the set of remaining tweets by applying Natural Language Pre-processing (NLP) steps. All the tweets are pre-processed by filtering out URLs, mentions, punctuations, and stop words. Then, we apply the lemmatization algorithm using the NLTK WordNet Lemmatizer[1], which considers the morphological analysis of the words and returns words to their lemmas. After processing the tweets, we extract proper nouns that appear in the tweets utilizing the Stanford Named Entity Recognition (NER) model[2]. We found that using the standard NER model has resulted in incorrect identification of Named entities due to the unstructured nature of the social media posts and the lack of proper spelling and grammar used in formal writing. Accordingly, we opt for utilizing the caseless model developed by the same research group in order to minimize the misclassification of any named entities included in the tweets.

After preparing the tweets, we extract a unique set of bi-gram words that appear in tweets posted during the course of the detected sub-event. We calculate the term frequency TF for *bi-gram* words with respect to time. In other words, we calculate the TF for terms within the time interval of the detected sub-event. Proper nouns usually have high probabilities of being related to the detected sub-events [121]; therefore, we assign more weight to the calculated TF of proper noun terms using a weight boosting factor x. Based on the literature [121, 109], the value of x is equal to 1.5 if the terms has been identified as a proper noun otherwise $x = 1$. Thus, the weight of w is calculated as $tf_w * x$. This approach in extracting representative terms will lead to assigning more weight to common terms that are present over the course of the entire event and not associated with a specific sub-event. In order to extract new significant terms that are specific to a certain time window regarding a selected sub-event, we measure the frequency of a term w within the detected spike at time t_i, compared to how common the term was within the preceding spike at time, $t_i - 1$, as shown in equation (6.2). This approach allows us to decrease the weight assigned to general terms and appoint higher weight to specific keywords that are related to the current sub-event. Finally, we rank each *n-gram* based on its significance

[1]https://www.nltk.org/api/nltk.stem.html#module-nltk.stem.wordnet
[2]https://nlp.stanford.edu/software/CRF-NER.shtml

score (equation 6.2) and select the Top-N terms to describe the detected sub-event type.

$$\textit{Score-sig}(w_{t_i}) = (tf_{t_i}(w) * log\ \frac{tf_{t_i}(w)}{tf_{t_{i-1}}(w)}) * x \tag{6.2}$$

$$x = \begin{cases} 1.5 & if\ x \in PN \\ 1 & if\ x \notin PN \end{cases}$$

where tf is the term frequency of the key word w, x is the boosting factor, and PN is the set of proper nouns.

When significant sub-events occur, fan discussion regarding the sub-event continues long after the time frame it happened [138, 94], which could lead to falsely identifying the same sub-event as a new incident. Accordingly, we introduce a merging step for combining spikes that reference identical sub-events in order to create one unified sub-event. For example, in the 2018 FIFA World Cup final match between France and Croatia; the Croatian player scored an own-goal which gave the lead to France. This sub-event generated a high volume of relevant tweets for a period of a few minutes. The incident (own-goal) is labeled a sub-event the moment the first spike in relevant tweets is flagged. Then, in the following minutes, if we detect a new spike which indicates discussion of the same own-goal sub-event, we then merge it with the same previously detected sub-event without signaling an occurrence of a new sub-event. To determine if two spikes are discussing the same sub-event, we compute the similarity between the vocabulary sets appearing within the spans of the two spikes. We utilize the cosine-based similarity metric to calculate the similarity score between the vocabulary vectors of the current detected spike and the previous spike. The cosine-based similarity takes two vectors of vocabulary that appear within the spikes' time windows, S_i and S_j, and quantifies their similarity according to their angle, as shown in equation (6.3). If the similarity score is greater than the threshold θ then the two spikes are considered as similar sub-events and merge to represent one unified sub-event. Otherwise, the sub-event detection method will signal an occurrence of a new sub-event. The threshold θ is defined based on historical experiments. The pseudocode of sub-event detection and identification process is demonstrated in Algorithm 1.

$$similarity(S_i, S_j) = \cos(S_i, S_j) = \frac{\mathbf{S_i} \cdot \mathbf{S_j}}{\|\mathbf{S_i}\|\|\mathbf{S_j}\|} \tag{6.3}$$

Algorithm 1 Sub-event Detection and Identification Algorithm

procedure SUB-EVENT DETECTION(F, α, β)

Input: Posts frequency per time window $F = \{f_1, f_2, ..., f_i\}$, T: List of posts in time i ;

    while recieving stream of data **do**
        $mean_n = mean(f_1, f_2, ..., f_i)$
        $std_n = std(f_1, f_2, ..., f_i)$
        $DTS_i = \alpha * (mean_i + (\beta * std_i))$ ▷ DTS: threshold for spike frequency
        if $F_{(t\_i)} > threshold$ **then**
            T=clean-preprocess(List of tweets)
            PN = *extract-named-entity(T)*
            Vocab = extract-bow(T) ▷ Vocab is a set of unique Bi-gram
            for all $v \in$ Vocab **do**
                $score_{v_{t_i}}$ = S-sig(v_{t_i}) ▷ calculating the significance score for each term v
            end for ▷ in the vocabulary set
            sim-score = semantic-sim(V_i, V_{i-1}) ▷ V_i is vocab set in time t_i, V_{i-1} in t_{i-1}
            if $simscore < \theta$ **then**
                Signal: NEW SUB-EVENT
                Top-N ← Rank-Bigram(Vocab) ▷ rank the list of terms based on their
            else ▷ significance score
                Merge current sub-event with the previous
            end if
        else
            No Spike Detected, Continue
        end if
    end while

end procedure

6.3 Tweets Categorization

In the previous section, we shed light on the sub-event detection process and the sub-event representative keyword extraction. We now look at analyzing the sentiments of each team's fans and producing a subjective summarization based on fan reactions. In order to do this, we need to categorize the teams' tweets. Assigning tweets to a particular team is not a trivial task. Fans, during a football match, tag their tweets in many ways, including the match hashtag (#FRACRO, #CROFRA), hashtags for multiple teams at once (#FRA, #CRO), or individual team hashtags. Tang and Boring, in [132], identified tweets with only one team hashtags as tweets written by this specific team's fans. Depending only on hashtags could result in an incorrect assignment, as fans are also using their opponent team hashtags when commenting on their way of playing or mocking the players' performances [132]. To identify specific teams' tweets, we propose a hybrid approach that consists of two phases: a hashtag-based step and a user-profile step.

First, we assign tweets to a team based on the inclusion of the team's hashtags, or specific keywords that only belong to that particular team. Team specific keywords would, in our opinion, include all the official team hashtags, nicknames, abbreviations, and player names. In this step, tweets are categorized into three sets: tweets belonging to team A, tweets belonging to team B, and mixed tweets that contain both teams' keywords. Categorizing tweets using only keywords or hashtags is not a perfect way of tweet separation. To further refine the set of tweets assigned to each team, we leverage the fans' public social profiles for determining their preferred teams. Fans usually tweet about their favorite team using specific keywords related to their team more than other teams over a series of matches [138, 33]. To ensure that each tweet assigned to a team is written by the team fan, we extract all the tweets written by a particular user during a set of previous matches. We then calculate the number of times the user mentions the team using the team's official hashtags, Twitter accounts, abbreviations, nicknames, and players' names. We determine users who have more than five tweets related to football matches worth considering for our purposes. Then, the list of the teams is ranked based on the calculated frequency score, and the top-N teams are selected. We restrict our ranking to the top-N teams to eliminate teams that have low-frequency scores. For a specific match, we characterize a user as a fan of team A if the difference between the team A frequency score and team B is greater than four.

6.4 Sentiment Analysis

Fan tweets reflect the sequential changes in fan sentiment through their reactions to sub-events occurring throughout the match. We track the changes in fan sentiment by performing sentiment analysis on the tweets belonging to each team's fans, which are extracted using the tweets categorization module. We utilize our sentiment classifier that is designed for analyzing sentiment expressed in football-specific tweets and described in chapter 5 to interpret the sentiment. The SVM-based model is employed for analyzing the sentiment of fans due to its consistent performance over different experimental settings. To identify the sentiment expressed in the fan's tweets, we extract the same features used to build the model which are discussed in detail in chapter 5. We then aggregate the sentiment of each team's fans to visualize the change in their feelings when a sub-event occurs.

6.5 Visual Content Filtering

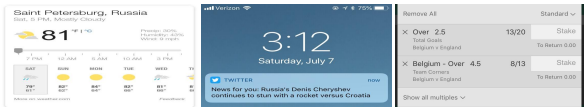
During an event, users in social media post all types of images using event hashtags to promote their content, often without consideration of their images' relevance to the event [1]. A large portion of the images posted on social media includes screenshots, adverts, memes, and reaction images that need to be filtered before generating the visual summary of an event [90]. Examples of different types of images extracted from social media during the FIFA World Cup 2018 are illustrated in Figure 6.1. Most of the previous objective summarization works excluded memes and reaction images, deeming them as inappropriate for the summary even though they may have been relevant to the event [92, 121]. Our work aims to produce a subjective summary that encapsulates public reactions and opinions during the event. Memes and reaction images consider as a tool to express fans' opinions and feelings about major sub-events occurring during the match [80]. Fans use them to praise their team or ridicule the opponent team [80]. Accordingly, in our work, we consider these memes and reactionary images in the generated summary to represent fan reactions throughout the event. We filter out screenshot images, diagrams, or maps deeming them inappropriate for subjective event summarization [14].

To ensure the relevance of the visual content, we first employ a rule-based filtering schema to filter out advertisements, thumbnails and small images. Images with a width or height less than 200 pixels are ignored as they are too small to be included in the summary. Likewise, images with dimensions that match the standard web banner advertisement

(**a**) Real photos

(**b**) Memes

(**c**) Reaction photos



(**d**) Screenshots photos

(**e**) Computer generated photos

(**f**) Diagrams photos

(**g**) Maps photos

Figure 6.1: Variation of images' types in social media content.

dimensions are excluded[3], as demonstrated in [92]. Following these heuristics is the first step in filtering out a portion of unsuitable images for summarization purposes. Images belonging to other categories are not recognized using the rule-based filter approach. Most of the image categories discussed (screenshots, memes, and diagrams) are synthetic images, some of which are considered inappropriate for event summarization; therefore, we classify the images into synthetic or real photo categories, making it easier to review and discard images that belong to specific types of synthetic categories which do not suit our purposes, such as screenshots and diagrams.

Some related works, such as [142, 82], have addressed the problem of synthetic image classification. These, however, are lacking in social media specific-image types such as screenshots, memes and reaction images. The NPIC system proposed in [142] classified synthetic images into five distinct classes: maps, icons, figures, art work, and cartoons. We adapt the synthetic images categorization of NPIC system [142] to include the missing classes of social media specific images as shown in Figure 6.2 and eliminate images categories that are not common within social media content such as art work [90]. To build

[3]https://commons.wikimedia.org/wiki/File:Standard_web_banner_ad_sizes.svg

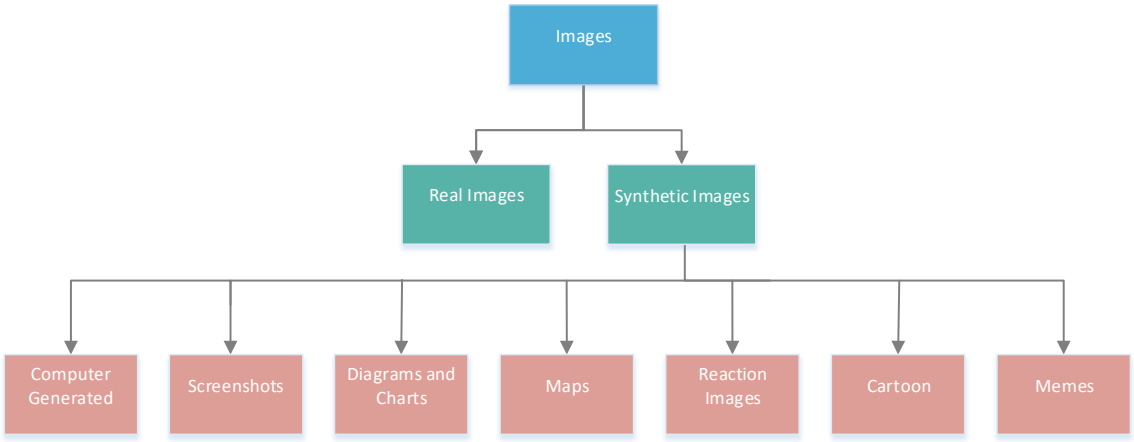


Figure 6.2: Images taxonomy for synthetic images considered in our work. Adapted from [142]

our classifier, we construct an image dataset containing various types of synthetics and real(natural) images, named Extended-NPIC which expand the NPIC to include the social media specific images and update the type of images included in the NPIC dataset. The details of the dataset will be described in section 6.8.1. For each image, we extract the following set of visual features:

- Color Histogram (CH): a representation of an image's color distribution. It represents the frequency of specific image pixels that belong to each color range. We extract a 3×256 dimensional histogram from the RGB color channels.
- Bag-of-Visual-Words (BoVW): the idea behind BoVW is adapted from the well-known Bag-of-Word (BoW) concept in information retrieval. It is based on representing an image as a set of features using SIFT descriptors. To generate the BoVW feature, we consider a 2-layer spatial pyramid and max pooling technique, as utilized in [17].
- Local Binary Pattern (LBP): a popular descriptor that computes a local representation of texture [52]. This descriptor is constructed by comparing the value of each pixel with its neighbor set of pixels. The feature vector is generated using the uniform variation of the LBP descriptor proposed by Ojala et al. [105], which results in 59-dimensional feature vector.

- Deep learning (FC7): we adopted the pre-trained Convolutional Neural Network (CNN) AlexNet Caffe model, [75], to extract the deep semantic features. Caffe implementation of the CNN model consists of five convolutional layers followed by three fully connected layers, trained on the ImageNet dataset. More details of the CNN architecture can be found in [63]. We extract features from the last fully connected layer (FC7) which are represented as 4096-dimensional feature vector.

After extracting different types of features, we utilize an SVM learning algorithm to build a classification model that is able to classify a specific image into one of eight classes (real photos, memes, reactions, cartoons, screenshots, diagrams, maps, computer-generated images). Images that are classified as screenshots, maps, or diagrams are discarded while other images are passed to the popularity prediction model.

6.6 Images Popularity Prediction

The refined set of images posted during the sub-event time is then passed to the popularity prediction model in order to predict its popularity ranking score. The development of the image popularity model is discussed in chapter 4. Our prediction model is designed for predicting the popularity ranking score of general set of images that are not associated with a specific topic or event. The proposed model can be adapted to predict event-specific images from various social media platforms by considering the difference in non-visual features. In order to investigate the capability of our model in predicting the ranking score of Twitter images, we test the model on a set of 25,000 Twitter images sampled from the FIFA 2018 World Cup event. Due to the variation in Flickr and Twitter contextual and textual features, we utilize a version of the model that is trained on visual features only. Our model has achieved a 0.20 rank correlation score when predicting the number of retweets and likes of the test images. We first employ our model to predict the ranking score of images that are associated with tweets posted during a sub-event. Then, the type of each image and its popularity score are sent to the summary generator module to generate the visual summarization of the specific football match.

6.7 Visual-Textual Summary Generation

The summary generation component is activated when a new sub-event is detected. The summary generator is populated with a set of tweets and images that correspond to the

detected sub-event. The summary generator selects the most representative text and images to describe the identified sub-event. Then, the textual and visual representation are aggregated to portray a subjective multimedia summary of the event based on fans' perspectives.

6.7.1 Textual Summary

This component is responsible for selecting a set of representative tweets that describe the detected sub-event. It selects the top ranked tweets from the set of tweets that is posted within the time window of the detected sub-event. Before ranking the tweets, we filter out text that includes more than four hashtags, three mentions, or two URLs. Unlike previous work, we do not eliminate short text that contain less than three words as in [13], or six words as applied in [121]. We focus on presenting subjective description of the event, and showing how fans behave during the time of the match. Fans cheer for their preferred team using colloquial language ("Cheer England ", "goooooaaaal", "hazard scores!#bel"), or they may simply acknowledge the game status ("halftime 1-0", "KICK OFF! #BEL #ENG"), or express anger at their team's performance, or frustration at the referees ("poor #eng", "Neymar= Disgrace!! #WorldCup", "Thanks referee #BEL #BRA"). We keep this type of text as it illustrates the reaction of fans towards a sub-event and can provide an insight into the type of the sub-event that has occurred. We select representative tweets by calculating the score of each tweet based on the sum of its terms' weight. The weight of each term is based on term frequency TF during the time of detected sub-event multiplied by a boosting factor if the terms is identified as a proper noun or a player name. Then the Top-N ranked tweets are selected for each team as a descriptive text for the detected sub-event. The number of tweets is decided based on the application and the user's preferences.

6.7.2 Visual Summary

For visual summarization, we rely on the popularity ranking score that we predicted using our developed model for selecting the most popular images. We select the top-N ranked images based on popularity. These images fall into one of the following categories: real photos from the pitch, real photos from outside the pitch, memes, reaction images, or computer generated images. For each team, the fan reactions are visualized using images that are associated with tweets posted by the team's fans.

6.8 Experiments and Evaluation

In this section, we describe the datasets used in our experiments. We then evaluate the performance of the main components in our subjective multimedia summarization framework: sub-event detector, image-types classifier and the generated summary.

6.8.1 Datasets

We use the following two datasets to evaluate the different components in our framework:

- The 2018 FIFA World Cup dataset: To evaluate our approach, we collect public tweets from Twitter posted during the 2018 FIFA World Cup competition using the Twitter Standard Streaming API [4]. The data is collected from June 14 until July 15 2018 using predefined list of keywords. The list of keywords includes the competition official hashtags: #FIFAWorldCup18, #WorldCup18, #WorldCup2018 and #WorldCupRussia2018. To enrich our dataset and ensure the coverage of varying matches during the 2018 World Cup, we used match-specific hashtags such as: #SUISWE, #ENGCRO and #CROFRA. We collected English tweets combined with their embedded images, if available, and all the metadata associated with the tweets such as: tweet identifier, time stamp, user information, and location. In our experiments, we consider matches within the round of 16, quarter-finals, semi-finals, third place and the final match due to make up for missing tweets in the previous stage. Each match is considered a standalone event. The total number of tweets collected during the 13 different selected matches is 987,528. 105,666 of these tweets contain images posted by 153,703 users. The selected matches are chosen based on the availability of data for the match playing time, and the burden of the annotation process. To ensure data quality, we filter out retweets, duplicated tweets, and tweets consisting of hashtags and/or mentions only.

- Extended-NPIC dataset: To identify image types, we use a dataset that consists of various synthetic and real images to train our model. To the best of our knowledge, the NPIC is the only publicly available dataset for synthetic and natural image classification introduced by Wang and Kan [142]. It includes images belonging to cartoons, figures, icons, and maps categories which are collected using class-specific keywords in order to retrieve a noisy labeled set of images via the Google image search

[4]https://developer.twitter.com/en/docs/tweets/filter-realtime/overview

engine. The NPIC dataset, however, lacks images that are common in social media, such as memes, reaction images, and screenshots. To overcome the limitation of the NPIC dataset, we collected images that reflect the type of images shared among social media users through Twitter and the Google search engine. We used the following keywords for the image collection process: memes, sports memes, cartoons, reaction images, maps, diagrams, charts, and screenshots. Our custom collection of images consists of 10,107 images manually sampled and belonging to synthetic and natural image categories. To train the classifier, we randomly select 70% of the real and synthetic photos as a training set, and the remaining images are used for testing.

6.8.2 Sub-Events Ground Truth

To annotate the FIFA World Cup 2018 dataset used in our experiments, we take advantage of sporting websites that provide live coverage of the matches during football competitions. We consider two websites: ESPN.com[5] and News18.com[6] as references when collecting the sub-events of each match and their descriptions. For the purpose of evaluation, we consider eight key sub-events utilized in the previous works [101, 94]: goal, own-goal, red and yellow cards, the match start time, halftime, match end time, and penalties. We manually label the output of the event detection algorithm by the type of sub-event in comparison to the reference description. Detected sub-events that belong to different categories of football sub-events such as foul, injury, or missed attempt, are excluded in the evaluation of the sub-events detection algorithm. Other detected sub-events are considered noisy signals.

6.8.3 Evaluation and Results

In this section, we describe our evaluation metrics, and performance results of each component including quantitative and qualitative results.

6.8.3.1 Evaluation of Images Types Classification

To evaluate the performance of the image type classifier, we use the accuracy, precision, recall, and F-score metrics. The evaluation metrics are defined in Chapter 5, Section 5.4.1.

[5]https://www.espn.com/
[6]https://www.news18.com/fifa-world-cup-2018/

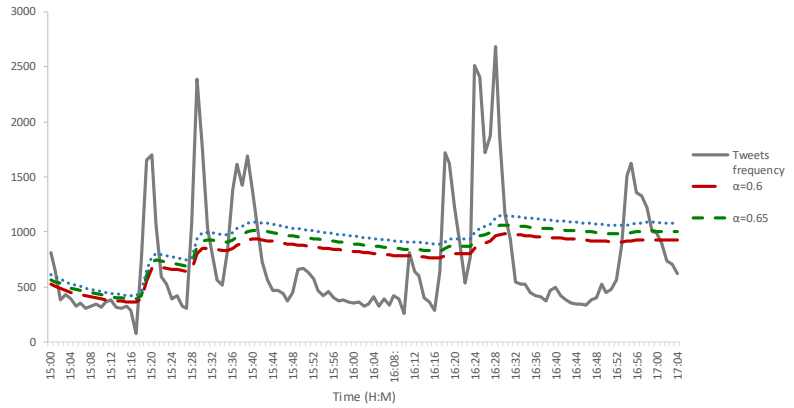
Table 6.1: Performance results of Linear and Kernel SVM using different features in identifying the type of an image

Classifier	Features	Accuracy	Precision	Recall	F-score
Linear SVM	FC7	0.82	0.79	0.76	0.77
	BoVW	0.73	0.75	0.63	0.65
	Color	0.54	0.49	0.41	0.38
	LBP	0.50	0.41	0.34	0.30
	All features	0.83	0.81	0.78	0.80
Kernel SVM	FC7	0.83	0.81	0.79	0.80
	BoVW	0.78	0.76	0.73	0.74
	Color	0.60	0.51	0.49	0.49
	LBP	0.57	0.49	0.44	0.43
	All features	**0.84**	**0.82**	**0.80**	**0.80**

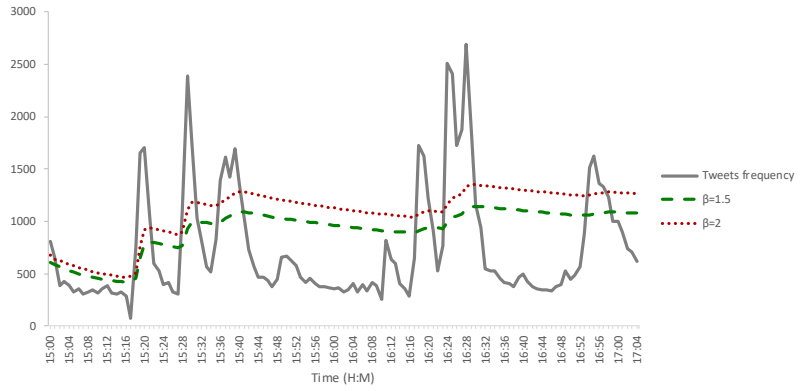
Table 6.1 reports the performance results of linear and kernel (RBF) SVM algorithms in classifying images into one of the eight classes considered in our work. For both classifiers, the results show that the deep learning FC7 feature outperforms all other hand-crafted features. Among the low level features, BoVW obtained the best performance at 73% accuracy using linear SVM. This accuracy increases to 78% using Kernel SVM. The results presented in Table 6.1 clearly show that the selected kernel SVM algorithm achieves better performance in comparison to the linear SVM. By analyzing the classification results, we find that the memes class is one of the more difficult image categories for the model to recognize, due to the fact that memes have diverse captioned text in conjunction with the same background images.

6.8.3.2 Evaluation of Sub-event Detection

In this section, we evaluate the performance of the sub-event detection algorithm. We compare the performance of the algorithm to the ground truth obtained from the references as described in section 6.8.2. We use recall and F-score as our evaluation metrics. Recall is the ratio of the number of key sub-events detected by the algorithm to the total number of key sub-events in the matches. F-score is the harmonic mean of recall and precision (the ratio of the number of key events detected by the algorithm to the sum of the number of key events detected by the algorithm plus the noisy events).



(**a**)Change in the threshold value with respect to different α values. β value is fixed to 1.5



(**b**)Effect of different β values on the threshold where α is equal to 0.7

Figure 6.3: Threshold values change when using different values for α and β parameters

6.8.3.3 Effect of parameters

Here we provide empirical results that highlight the impact of the algorithm's parameter adjustments. The sub-event detection algorithm has three parameters that can be set to a range of values: the decay factor α, the β coefficient and the semantic similarity threshold θ. To evaluate the effects of these parameters on the results, we assign different values to them as follows: $\alpha = [0.6, 0.65, 0.7]$, $\beta = [1.5, 2]$, and $\theta = [0.25, 0.30, 0.35, 0.40]$. The reported result is the average over the thirteen matches that are considered in our experiments.

The results in Table 6.2 show the overall performance of the detection algorithm in terms of recall and F-score for α, β, and θ parameters. We examined the performance of

Table 6.2: Performance results of the sub-events detection algorithm using different values for the parameters α, β, and θ

	α=0.6 and β=1.5		**α=0.6 and β=2**		**α=0.65 and β=1.5**		**α=0.65 and β=2**		**α=0.7 and β=1.5**		**α=0.7 and β=2**	
θ values	Recall	F-score	Recall	F-score	Recall	F-score	Recall	F-score	Recall	F-score	Recall	F-score
θ=0.25	0.78	0.45	0.71	0.47	0.74	0.48	0.65	0.51	0.65	0.50	0.63	0.56
θ=0.30	0.80	0.46	0.72	0.49	0.75	0.50	0.71	0.53	0.70	0.52	0.64	0.57
θ=0.35	0.80	0.44	0.71	0.47	0.74	0.47	0.69	0.52	0.70	0.51	0.64	0.55
θ=0.40	0.80	0.44	0.71	0.46	0.74	0.47	0.69	0.52	0.70	0.51	0.66	0.56

the detection algorithm with respect to increasing the value of α parameter. Assigning a small value to α will generate a set of a large number of spikes during the event time. On the other hand, increasing the values of α will increase the threshold value and reduce the number of detected spikes, as depicted in Figure 6.3. We compare the results of assigning different values to α, and consider fixing the other two parameters $\beta = 1.5$ and $\theta = 0.25$, where these values range from 0.6 to 0.7. From the results in Table 6.2, we can see that when $\alpha = 0.6$ we have higher recall which means detecting more relevant sub-events while larger values decrease the recall percentage. Yet, assigning small values to α will result in detecting more non-sub-event spikes which affect the performance of the algorithm. Similarly, increasing the value of β will increase the threshold value and result in generating smaller number of detected spikes yet reduces the noisy spike. When comparing the performance of the θ parameter, we observe that $\theta = 0.35$ and $\theta = 0.40$ have similar performance in terms of recall. From the results, we can see that $\theta = 0.25$ has the lowest recall score in general. Assigning small value to θ means that the similarity score between the vocabulary appearing in the two spikes should be small, which means having very divers words in the two spikes. However, in sub-event detection task, common words will appear more frequently due to the similarity of the event context. The results show that assigning θ to 0.30 achieves the best performance in terms of recall and F-score.

Table 6.3 illustrates the performance of the detection algorithm while identifying individual key sub-events that are considered in the evaluation. We can see in the results that the algorithm is able to detect most of the goals, own-goals, penalties, and red cards. Our algorithm missed three goals in three different matches: Spain vs Russia, Columbia vs England, and Brazil vs Belgium. In the match between Spain and Russia, the goal was scored after a penalty shot and the frequency of the tweets were high during the minutes of the penalty shot sub-event which affected the detection of the goal. Similarly, in Columbia vs England match, fans were discussing the penalty more actively than the goal, and also, the goal was reflected late in their tweets, which lead to the missing of this sub-event. The algorithm only detected 17 yellow card sub-events out of the 47 which occurred during the

Table 6.3: Recall of individual key sub-event detection

Event type	# of actual event	Recall
Goal	36	0.92
Own-Goal	3	0.67
Red Card	1	1
Yellow Card	47	0.36
Penalty	4	1
Match Start	13	1
Half Time	13	0.77
Match End	13	0.85

considered matches in the evaluation dataset. Because yellow cards are not as significant or impactful as red cards, penalties or goals, fans are less likely to comment on this type of sub-event. While investigating the reasoning behind the missing yellow card detections, we found out that, in some cases, the fan tweets made no mention of the occurrence of the yellow card sub-event during the match. In other cases, because the yellow cards were given to the players near the match end time or at the same time as some other major sub-event, such as a penalty, fans were eager to discuss the more significant sub-event instead of commenting on the yellow card sub-event.

6.8.4 Evaluation of the Generated Summarization

Evaluating event summarization is a challenging task due to the subjectivity of the problem; therefore, in order to evaluate the summarization performance, we have used two experimental settings, as follows:

- Automatic summary evaluation: in this evaluation, we generate an objective summarization of all the tweets that are posted which represent the detected sub-event. Then, we employ the *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE-N) metric [83], which is widely utilized to measure textual summarization performance, to compare our automatically generated summarization to the reference summary created by humans for sport websites such as ESPN.com. The calculation of ROUGE-N metric is shown in equation (6.4).

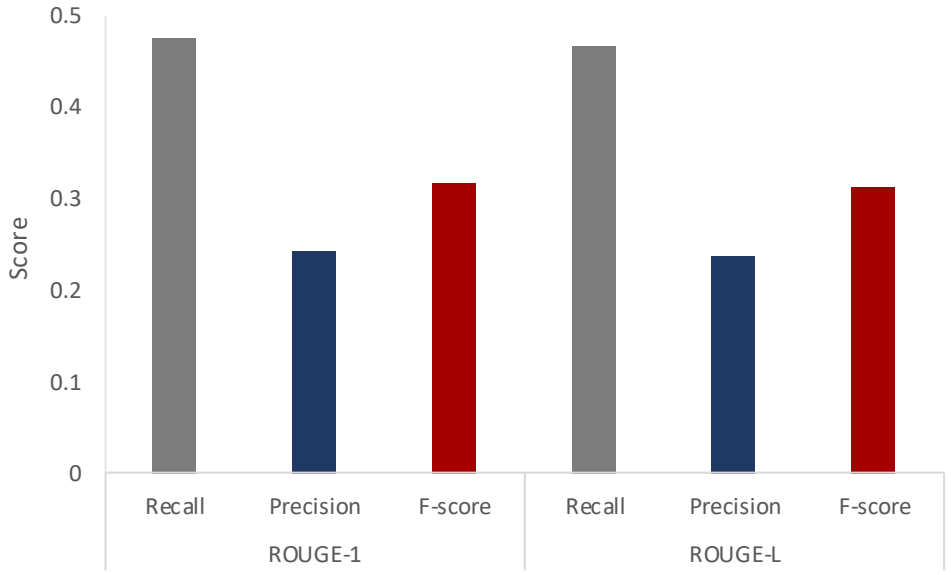


Figure 6.4: Performance results of our summarization algorithm compared to recap article as the reference summary.

$$ROUGE\text{-}N = \frac{\sum_{S \in RS} \sum_{N\_gram \in S} Count_{match}(N\_gram)}{\sum_{S \in RS} \sum_{N\_gram \in S} Count(N\_gram)} \tag{6.4}$$

where N_gram and $count_{match}(N\_gram)$ is the maximum number of n-gram co-occurring in the automatically generated summary and a reference summary.

- User study-based evaluation: we conduct a user study to evaluate the effectiveness of our generated subjective summary through an online questionnaire. The purpose of the study is to evaluate the coverage of the most important sub-events which transpired during the selected matches. Also, we evaluate if our summary contains the same information in comparison to the reference summary provided by the sport website for detected sub-events. In addition, we evaluate if the generated summary illustrates the two views of the same event based on the fans' perspectives (show subjectivity in describing the same event).

6.8.4.1 Automatic Summary Evaluation

We evaluate the performance of our summarization algorithm compared to the human-generated summary on the ESPN.com website using ROUGE-1 and ROUGE-L for the evaluation. In comparison, ROUGE-1 uses 1-gram words to relate the automatically generated summary to the reference summary, while ROUGE-L uses the longest common subsequence between the two summaries [83]. Figure 6.4 shows the ROUGE-1 and ROUGE-L results when comparing our summary to the reference summary using a recap article. The reported results represent the average from the thirteen matches that are used for the performance evaluation. The automatically generated summary achieved 48% in terms of recall and 32% in terms of F-measure using ROUGE-1. In this type of evaluation, we expect to see a weaker performance compared to human-based summarization due to the differences in the vocabulary used. We are comparing summaries generated from tweets to journalistic types of descriptions. Tweets are full of emotional words and slangs, whereas this is not the case in a journalistic article. Also, recap articles use an objective voice when describing the event, where tweets are completely subjective. We still find that our summarization achieved good results utilizing tweets. Table 6.4 shows a sample of the summary generated by our approach and the one obtained from the sport website for the third-place playoff match between Belgium and England.

Table 6.4: Sample of our summarization and the commentary article of ESPN.com for the third-place playoff between Belgium and England

Event	Our Summary	Recap Article
Kick off	kickoff!!! belgium 0 england 0 belgium get the game underway. #beleng #belgiumengland #worldcup kick off! third place play-off #eng vs #bel the beaten semi-finalist are battling it out for third-place	First Half begins.
Goal	belgium take the lead! england falls asleep at the back! meunier gets the goal! belgium 1 - 0 england ! #world-cup #beleng 4' goal!! meunier scores early to give belgium the lead in st. petersburg. belgium 1-0 England	Goal! Belgium 1, England 0. Thomas Meunier (Belgium) right footed shot from the centre of the box to the centre of the goal. Assisted by Nacer Chadli with a cross.
Half-time	half time. england 0-1 belgium. belgium are 45 minutes from sealing third place.	First Half ends, Belgium 1, England 0.
Goal	goal! belgium 2-0 england eden hazard gets his goal and all but seals third place. goal: belgium 2 - 0 england. eden hazard finishes england off and probably phil jones' england career.	Goal! Belgium 2, England 0. Eden Hazard (Belgium) right footed shot from the centre of the box to the bottom left corner. Assisted by Kevin De Bruyne.

6.8.4.2 User Study-based Evaluation

Evaluating automatically generated summarizations is not an easy task when taking into consideration that our goal is to generate a subjective summary. We conducted an online survey to evaluate the generated subjective summarization by football fans. We chose the online survey approach as it presented an opportunity to reach football fans of various ages with a wide range of interest levels in football, it also helped to minimize bias in a controlled experimental setting. Our aim of conducting the user study is to evaluate the generated summary based on the following aspects:

1. Information coverage and summary understanding: To measure if the summary covers the different sub-events that happened during the match and includes most of the information appear in the reference summary. We also measure how easy it is to understand the generated summarization.

2. Relevance to the event: To measure how much of the selected text and images are related to the match and actually describe what happened during the game. Also, we attempt to measure if the type of presented images is representative for summarizing a football match, or fans' reaction to the event.

3. Subjective view of the event: Attempts to measure whether the summary provides two distinct views of the same events based on fan perspectives.

- **Survey Overview and Design:** To evaluate the football match summarizations generated by our system, we conduct an online survey using Google sites and forms. In this user study, we evaluate the summarizations generated for two matches from the FIFA World Cup 2018: the third-place playoff (Belgium vs. England) and the final match (France vs. Croatia). We design two websites utilizing Google Sites in order to cover the summarizations of the selected matches individually. For the survey questions, we use Google forms that are embedded within the designed websites for the questionnaire. We group questions that discuss the same evaluation aspect into one part to avoid participant confusion. On the main page, we describe the purpose of the study and provide the instructions that help the participants to complete the evaluation. We also provide the participants with two links to the matches to remind them of the events that occurred during the matches. The survey opens with a pre-evaluation part, then follows with three main parts that discuss the summarization performance. Each part can be done independently, and participants may complete

the survey in their own time and convenience. The survey consists of 33 questions in total: 28 (7-point) Likert scale questions, 4 multiple-choice questions, and 4 short answer questions. All the answers received from the participants are stored in the cloud, specifically, on Google Drive.

The participants first start with a pre-evaluation questionnaire, which collects demographic information. Questions are asked to gather information about the participants' level of interest in football and usage of social media for tracking matches status updates, as well as their opinion about what type of sub-events they consider to be major key events during football matches. Participants may then navigate to the first part of the survey, where our goal is to familiarize the participants with the concept of match summarization using social media content and evaluate the overall summary of the match. Here, we also wish to evaluate to what extent changes in fans sentiment is a direct reflection of the occurrence of sub-events during the match. The second part of the survey focuses on evaluating the subjective summarization of the given match. We formulate the questions to cover the aspect of differing views of the same event provided by the teams' fans based on their perspectives. We include questions related to the textual and visual part of the summary. In the third part, we want to capture participants' preferences regarding the number of tweets and images that should be included in the summary, which provide a clear description of the sub-events. We also ask the participants to evaluate their overall experience of using this type of summarization compared to the text-based summary. Participants are given the choice of either evaluating both matches or choosing one of the two. The links for the study websites are included in both the invitation email and the social media post. We use the term "events" instead of "sub-events" in the user study in order to avoid any confusion among the participants.

- **Participant Recruitment:** We distribute an email invitation for the evaluation study through our networks and also posted the invitation on Facebook. The invitation is distributed among public football groups. The participation in our study was restricted to people 15 years and older who watching football matches. Each participant receives a 15 CAD e-Gift card as compensation, of choice, from Amazon, Best Buy, or Tim Hortons.

- **Participant Dynamics:** A total of 25 male football fans participated in our user study. We discarded 8 of them due to quality issues and incomplete information. 15

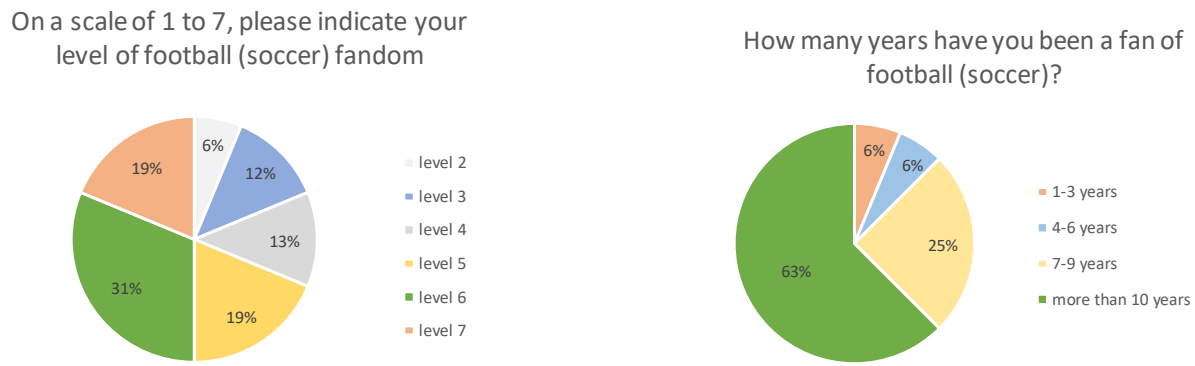


Figure 6.5: Participants levels of football interest and fandom.

of the 17 remaining participants evaluated the third-place playoff between Belgium and England, while 13 participants chose to evaluate the summary of the final match between France and Croatia with 11 of the participants evaluating both matches. 41% of the participants specified their age as between 20 and 29, 41% of them between the ages of 30 to 39, and 18% of the participants are between the ages of 15 to 19. 41% of the participants possess an undergraduate degree, 29% have a Master's degree, 18% of the participants are high school students, while the remaining hold a Ph.D. degree.

All the participants are interested in football and watched the 2018 FIFA World Cup event. As shown in Figure 6.5, 31% of the participants indicated their football fandom level as 6, while only 6% of the participants are at level 2. The majority of the participants have been football fans for more than ten years (63%), while 6% of the participants have been football fans for 1 to 3 years, see Figure 6.5. Eight of the participants are indicated that they follow football-specific accounts on social media, while only four indicated that they do not follow any football-specific accounts. Ten participants stated they always use social media for football match status updates, while three participants noted they only sometimes use social media for updates. The participants' responses show that there are different interest levels in regards to using social media for football event tracking. We asked the participants to select what they consider to be event of interest from a list of events that could happen during a football match. All the participants identified goals and penalties as major events in football matches. Red card and own-goal events also were selected by 12 and 10 participants as key events, respectively. On the other hand, events such as

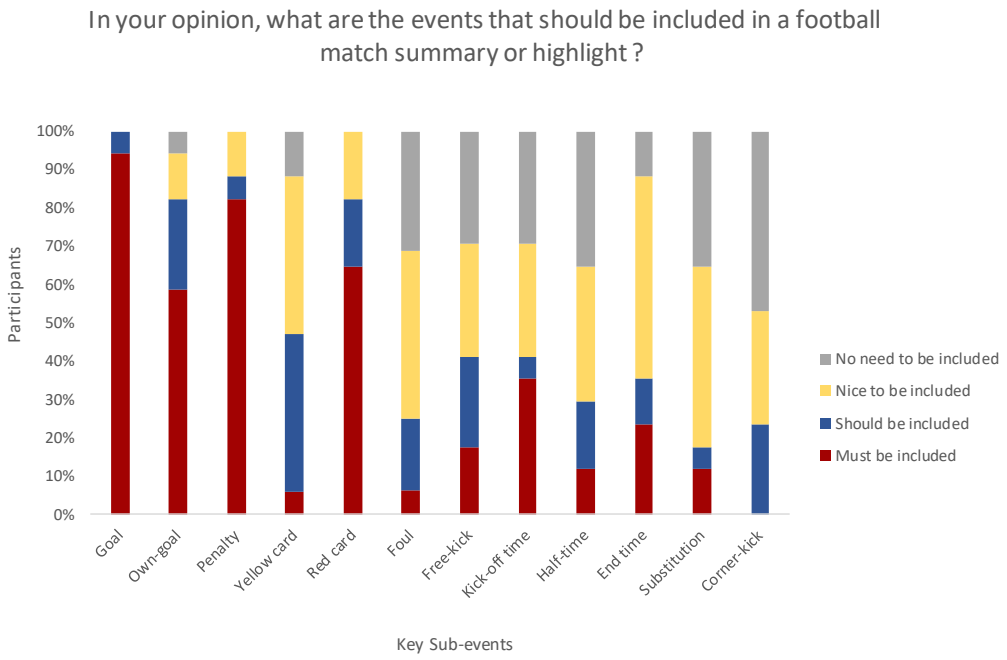


Figure 6.6: Sub-events and its importance levels to be included in the summarization of a football match as indicated by the participants

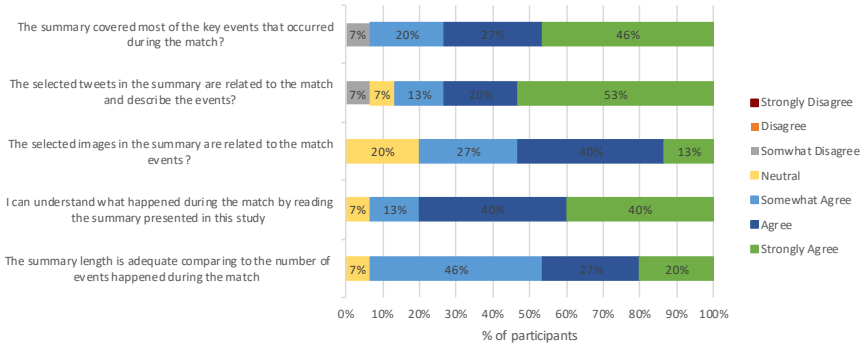
yellow cards, kick-off time, half-time, end time, and fouls were of less interest to the participants. When we asked the participants what key events must be included in a football match summary, 16 participants indicated that goals must be included in the summary, and 14 participants stated that the penalties must be included. Own-goal and red card events were flagged by 10 participants as a necessary inclusion for any football match summarization. Other events show less importance to fans, as shown in Figure 6.6. From the answers we obtained, we conclude that goals, penalties, red cards, and own-goals are key events that will affect fan satisfaction with a football game summarization.

6.8.4.3 Results and Discussion

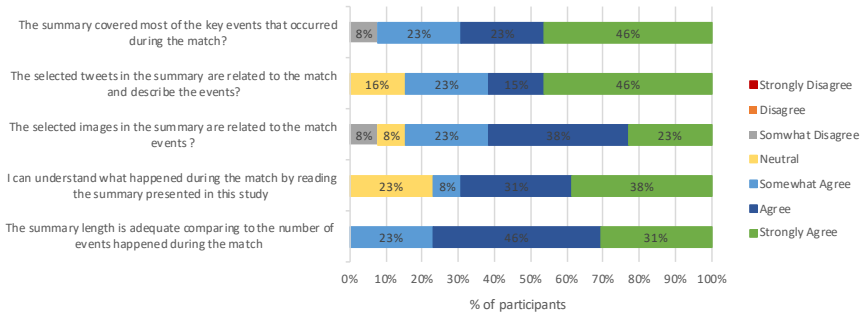
We analyzed the participants' answers to our survey regarding the summarizations of the two matches. Figure 6.7 depicts the Belgium vs. England match summarization presented to the participants. We asked if the summary in question covered most of the key events

Figure 6.7: Sample of the multimedia summarization that presented to the participants in our user study.

that occurred during the game in order to gauge if our summary adequately reflected the point of interest of the match. Seven of the participants who evaluated the third-place playoff strongly agree, and six of the participants who evaluated the final match also strongly agree. The median for responses to this question from both matches evaluation is 6, or, "agree" on a Likert scale. In regard to summary comprehension, the participants indicate they are able to decipher what happened during the match by reading the provided summary. Almost 77% of the participants who evaluated the final match agreed that the length of the provided summary was adequate compared to the number of events that happened during the match. To assess the relevance of the selected tweets and images in summarizing the matches, we asked the participants to which extent they agree or



(**a**)Belgium vs England match



(**b**)Croatia vs France match

Figure 6.8: Participants responses to different questions in regards to summarization coverage and representation of the given match

disagree that the presented tweets (images) are related to the match. Almost 73% of the participants who evaluated the third-place match showed an agreement, while only one participant disagreed with the statement. 8 participants agreed that the selected images are related to the third-place match, while 3 participants neither agree nor disagree. For the final match, almost 61% of the participants (8 out of 13) agreed that images presented in the summary are related to the match, while one participant somewhat disagreed. Figure 6.8 shows the percentage of participants who selected each Likert response to answer the above-mentioned questions. It is worth mentioning that images included in the summary belong to different categories. Some of the images show scenes taken live at the event itself, on the field, while other could be memes or computer-generated images. The differences between the image types may impact the participants' decisions about image relevance to the match. Therefore, we asked the participants which type of images they prefer in the objective summarization of a football match. Almost all participants included on-site images (real photos taken from the field during the event), and 10 of the participants also showed interest in seeing memes and reaction images included in the summary.

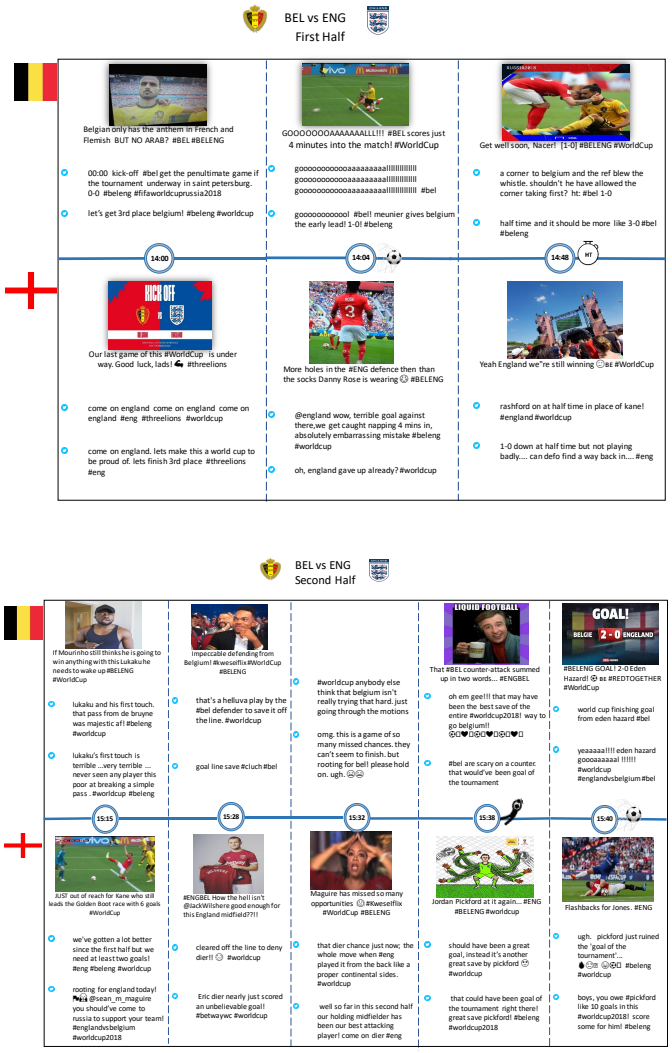
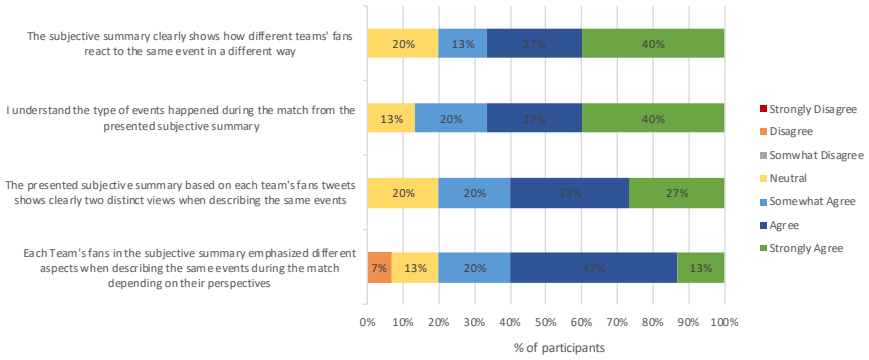
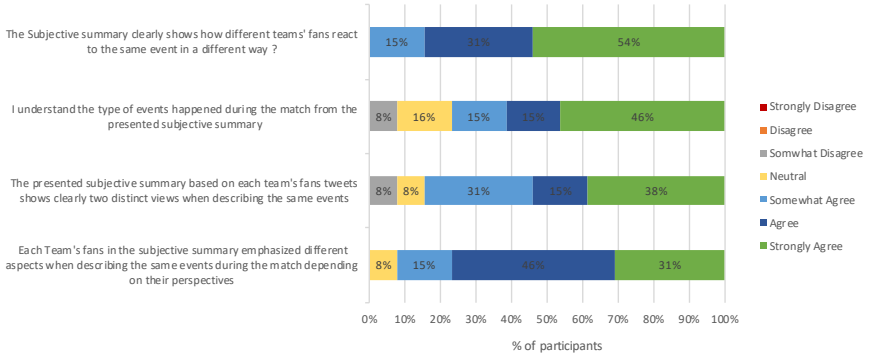


Figure 6.9: Sample of the subjective multimedia summarization that presented to the participants in our user study.

The main goal of conducting the user study is to evaluate the effectiveness of our approach in summarizing a football match based on fans' views. Sports, in general, are experienced differently by each individual fan, and they put emphasis on different aspects of the match based on their preferences. To balance this, in the second part, our questions are geared to determine to what extent our summary illustrates the two distinct views of the same events based on fans' perspectives. Consequently, we group different questions to evaluate this aspect in our subjective summarization for the two matches. First, we show the participants the generated subjective summarization of each match, Figure 6.9 depicts the subjective summarization of Belgium vs. England match. Then, we asked the participants to evaluate the following statement using 7-point Likert scale *" the subjective*



(**a**)Belgium vs England match



(**b**)Croatia vs France match

Figure 6.10: Participants responses in regards to the evaluation of the subjective summarization

summary clearly shows how different teams' fans react to the same event". Almost 67% of the participants who evaluated the third-place match agreed with the statement, and all the participants who evaluated the final match agreed. None of the participants disagreed with the stated statement in the question.

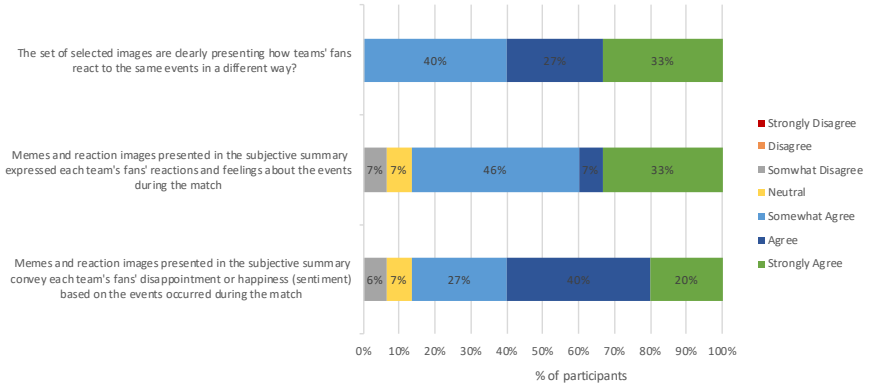
Another question about the same aspect is worded as follows, *"Does the presented subjective summary based on each team's fans tweets clearly show two distinct views when describing the same events"*. One participant selected the option somewhat disagree in the final match evaluation, while 53% of the participants (7 out of 13) agreed with the statement. For the third-place match, almost 60% of the participants (9 out of 15) agreed with the statement, while 3 participants neither agree nor disagree. Evaluating the summarization is subjective and dependent on the participants' opinions and point of view. Sport fans, in general, watch highlight summarizations that simply describe what hap-

Table 6.5: Examples of different reactions by Belgium and England Fans towards the same events

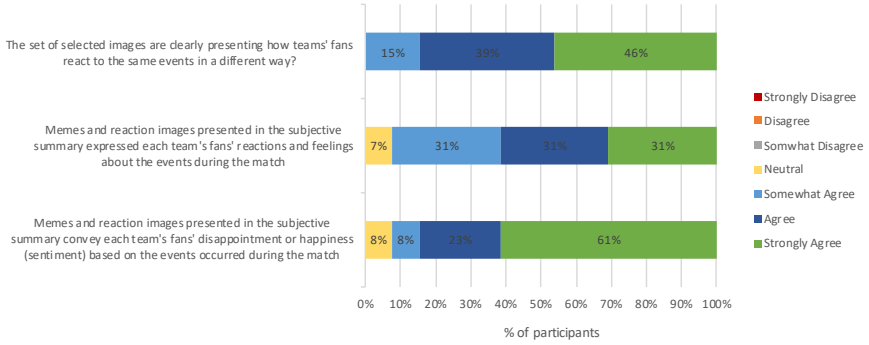
Sub-Event	BEL Fans Reaction	ENG Fans Reaction
Kick-Off	00:00 kick-off #bel get the penultimate game if the tournament underway in saint petersburg. 0-0 #beleng #fifaworldcuprussia2018	come on england. lets make this a world cup to be proud of. lets finish 3rd place #threelions #eng
Goal	goooooooooool #bel! meunier gives belgium the early lead! 1-0! #beleng	wow, terrible goal against there,we get caught napping 4 mins in, absolutely embarrassing mistake #beleng #worldcup
Half-Time	a corner to belgium and the ref blew the whistle. shouldn't he have allowed the corner taking first? ht: #bel 1-0	1-0 down at half time but not playing badly...can defo find a way back in..#eng
Attempt Saved	belgium is the best counter-attacking team at this #worldcup. simply sensational!	wshould have been a great goal, instead its another great save by pickford #worldcup
Goal	yeaaaaa!!!! eden hazard goooaaaaaal !!!!!! #worldcup #englandvsbelgium #bel	2-0. game over for england. we're finishing 4th. #engbel #eng

pened during the match. In contrast, they use social media to see others' opinions about the events that occurred during the match and to interact with other fans, discussing the incidents. We also take into account that our participants different ages, and have different experiences, which results in diverse subjective responses and levels of understanding of the goal of the evaluation. Accordingly, for our last inquiry on this matter, we ask the participants to pick their Likert responses to the following statement *"Each Team's fans in the subjective summary emphasized different aspects when describing the same events during the match depending on their perspectives"*. Ten participants of the group who evaluated the final match agreed with the statement, and one participant chose a response that was neutral. In the second group, nine participants responded as agreeing to the statement, three selected somewhat agree, and one participant disagreed with the statement. Figure 6.10 shows participant responses to the statements discussed above. Table 6.5 illustrates the different reactions to the same events which occurred during the match by Belgium and England fans.

Because our generated summarization is focuses on depicting how fans react to different events that occurred during the match, we included various types of images that would be considered noisy in an objective summarization. To evaluate if memes and reaction images convey fan emotions during the match, we asked the participants to select their agreement level to the statement *" memes and reaction images presented in the subjective summarization convey each team's fans' disappointment or happiness"*. 84% of the participants who evaluated the final match agreed with this statement. For the third-place playoff, 60% of the participants agreed with the statement, and one participant disagreed with the statement. Also, we asked the participants to what extent they agree or disagree with the following statement using 7-point Likert scale *"memes and reaction images presented in the subjective summary expressed each team's fans' reactions and feelings about the happenings events during the match"*. 7% of the participants who evaluated the final



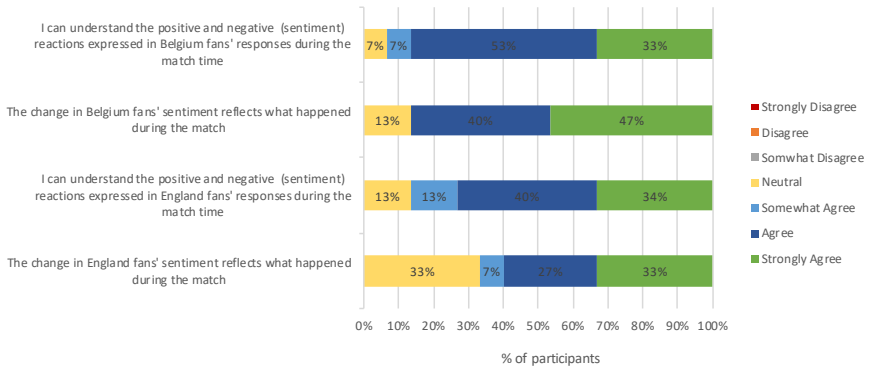
(**a**)Belgium vs England match



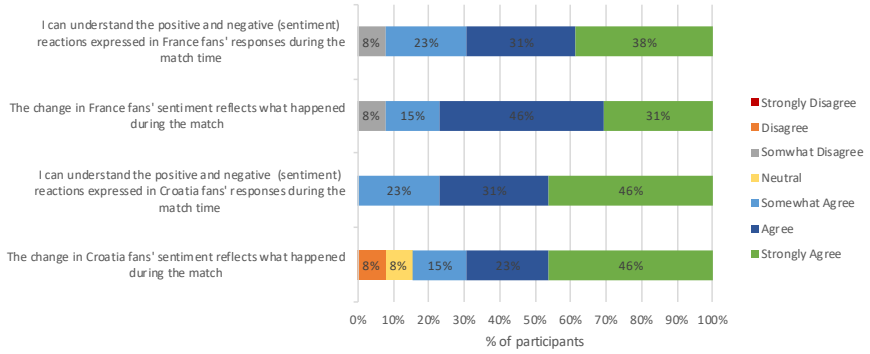
(**b**)Croatia vs France match

Figure 6.11: Evaluation of the subjective summarization- different stories by teams' fans

match selected the neutral option where they neither agree nor disagree. The remaining are equally distributed among somewhat agree, agree, and strongly agree with 31% for each Likert option. For the second group who evaluated the third-place match summarization, 33%, 7%, and 46% of the participants selected strongly agree, agree and somewhat agree, respectively. We asked the participants to identify the most suitable types of images that illustrate fans' opinions towards events which happened during football matches. Among the 17 participants, 13 selected memes and reaction images as the most suitable type of images, and 11 participants also included on-site images type. Note that the participants can select more than one option for this question. From the participants' responses, it is possible to say that memes and reaction images are similarly important in describing how fans react to events during sports match when compared with real-life images from the event itself. Figure 6.11 shows the percentage of participants who selected each Likert response to answer the above-mentioned questions.



(**a**)Belgium vs England match



(**b**)Croatia vs France match

Figure 6.12: Participants responses to different questions in regards to fans sentiment analysis during the given matches

To evaluate if the change in fan sentiment is correlated with the events that happened during the match, we asked the participants to select their responses to the following statements from strongly disagree to strongly agree on a 7-point Likert scale. First, we asked them if they understand the positive and negative reactions expressed in fans' responses during the match. The majority of the participants indicated an understanding of the sentiment expressed in fan tweets for both matches, as can be seen from the results depicted in Figure 6.12. Also, we asked them if the provided sentiment chart reflects the fans' sentiment changes during the match. Only one participant disagreed with the sentiment reflection for France fans, while 10 participants agreed. Regarding Croatia fans' sentiment, 9 participants agreed that the provided sentiment chart reflects the fans' sentiment change in a correlation to the events, while one participant showed a disagreement. For the Belgium and England game, five participants neither agree nor disagree with the sentiment reflection of England's fans. Thirteen participants agreed with the sentiment

reflection of Belgium's fans. The results are illustrated in Figure 6.12. In the match between England and Belgium, England's fans reported that their team missed many scoring chances; however, the sentiment reflected positive feelings when aggregated based on time. The reasoning behind this is the fans were happy about the England's performance in the second half compared to the first half of the match. Moreover, the fans generated a high volume of positive tweets when the goalkeeper made multiple saves against Belgium's offensive players.

To evaluate if the automatically generated summary includes similar content when compared to a manually generated summary, we show the participants the summary generated by the ESPN.com website of the same key events detected by our algorithm. We asked the participants to rate the generated summarization using a 7-point Likert scale based on inclusion of the information provided by the manual summary. Almost 69% of the participants who evaluated the final match agreed that our summary does convey the same information as the ESPN.com recap article for the detected key events. All the participants who evaluated the third-place match between Belgium and England agreed that the manual summary portrays similar information to the automatically generated summary except one participant who disagreed. Conclusively, we found that the automatically generated summary is able to provide similar information included in the manual summary of the matches regarding the detected events.

To identify the number of Top-N tweets and images that are suitable for summarizing a football match and delivering a comprehensive overview, we provide the participants with three versions of the summary. The first version includes one tweet and one image to describe each detected event that occurred during each of the matches. The second and third versions have two tweets-two images and three tweets-three images, respectively. The sample of the summaries with different numbers of top-N tweets and images is illustrated in Figure 6.13. We then asked the participants to indicate which version of the summary conveys the most effective description of the events. 62% of the participants who evaluated the final match summarization selected two tweets and two images as the summary that gave a good overview of the event. In the third-place playoff evaluation, 40% of the participants also selected two tweets and two images as the option that gave them a clear description of the events, while 40% selected one tweet and one image. See Figure 6.14. The common motivation among participants for selecting one tweet-one image summaries are "to the point information" and "easier to grasp while watching the match". On the other hand, participants who selected two tweets-two images summarizations to describe

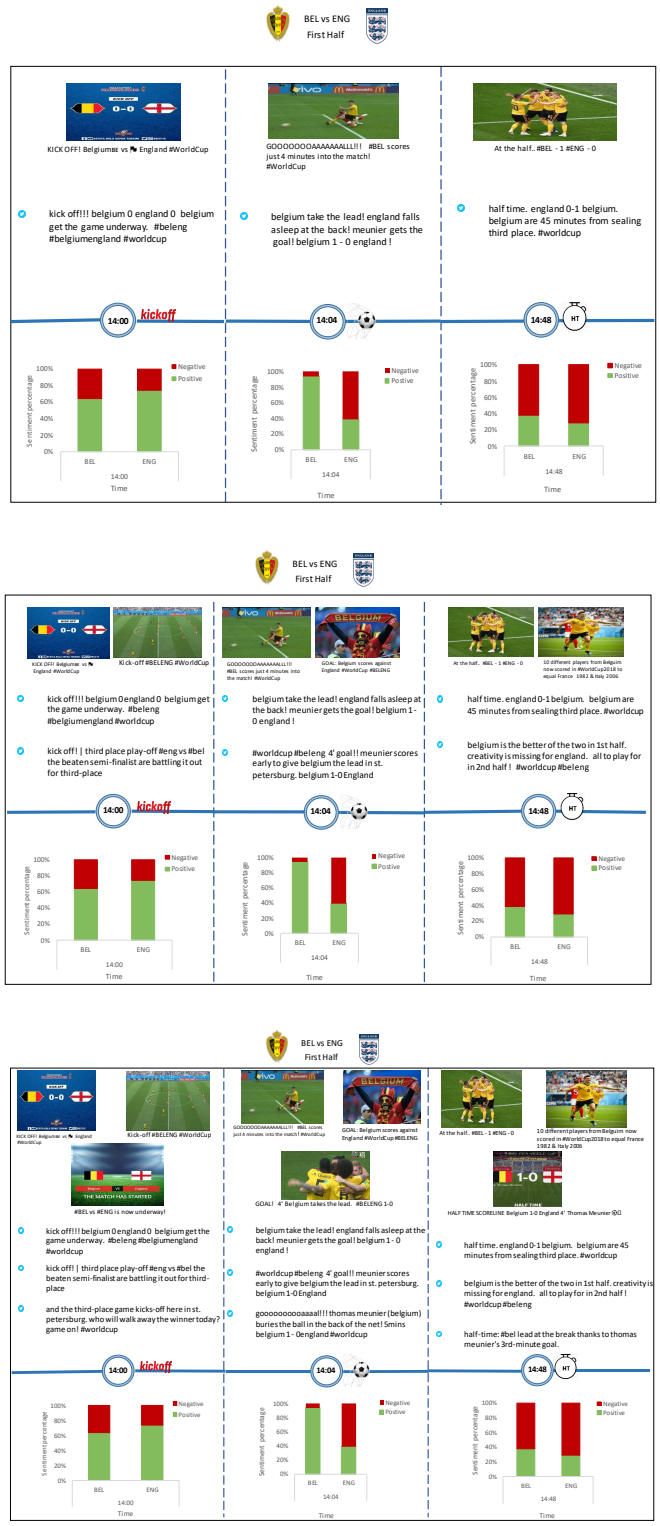
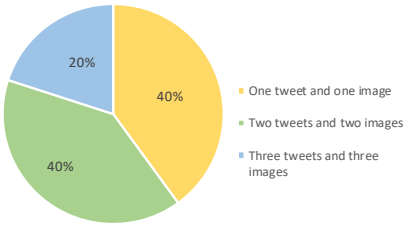
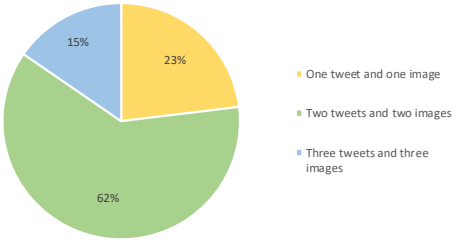


Figure 6.13: Different number of tweets and images included in the summary to describe what happened during the match.



(**a**)Belgium vs England match



(**b**)Croatia vs France match

Figure 6.14: participants responses in regards to the suitable number of tweets and images for summarizing a football match

the match said that they "included sufficient information", "has more details about the event", "gives a better overview of the event where they can compare the reactions of the teams' fans" and "feel the reliability of the provided summary".

At the end of the survey, we asked the participants to evaluate the following statement using 7-point Likert scale: *"I like to see a summary of fans' reactions during the match such as the one generated by our system"*. Six out of the seventeen participants in the survey said they strongly agreed with the statement, and seven participants picked the agree option from the Likert scale. On the other hand, two participants strongly disagreed with the statement. We asked the participants a follow-up question to provide a reason for their choices. One of the participants said that when he is watching the match live, he does not need fans' reactions. The other participant did not provide any reason for his choice. The participants who like the idea of summarizing fans reactions provide a wide range of reasons. For instance, they said the fans are an integral part of any match and they like to see the fan feedback. They also mentioned that when something particularly good or bad happens in a match, they like to see what other fans have to say about it. One of the participants mentioned that it is exciting to see fan reactions because they are often relatable to what he is feeling while watching the game. Some of the participants mentioned preference for summaries that contain video highlights; however, this is out of the scope of this current study.

Finally, participants were asked two open-ended questions: what you liked most about the presented match summarization, and, do you have any suggestions to improve our football match summarization? Most of the participants provided positive responses when answering the questions. The following is a sample of some of the answers we had from the

participants: "I like that it stated the important events during the match, such as the goal and the dangerous misses in the game", "short and concise information", "I like the mixed of tweets and images, makes it interesting and attention grabbing", "There are pictures of the events that happened and the design is very neat.", and "visualization , app design, fan reactions". For suggestions, participants mentioned adding videos to the summary and for it to support gif pictures. One participant recommended providing two options to the users- one short and to the point as described by the participant, and the other, a more detailed summarization.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

During real-world events, users generate and distribute timely updates of what is happening during the event using their social media accounts. This diverse content is rich with introspective insight, including the emotions, opinions and reactions of these users in direct correlation to how an event evolves. In many cases, having a handle on the state of mind of the public during a particular event, can be more useful than having simply an objective, chronological account of the who, what and where of said event. In this thesis, we address the limitations of creating event-based summarizations based solely on objective, textual data by proposing a new digital twin based approach which includes the consideration of subjective multimedia data while generating chronological summarizations of sporting events. The proposed sporting events digital twin summarization framework incorporates the image popularity prediction and sentiment analysis models in the process of generating the event-based summarization. Within a social stream related to a specific sporting event, we detect and identify the occurrence of important sub-events based on changes in the stream activities. We develop a sentiment analysis model in order to monitor the change in fans' sentiment as the event evolves over time. In the process of developing the sentiment model, we introduce a new sentiment dataset that designed specifically for the football-domain. Each instance in the dataset is manually annotated based on the multi-class sentiment. We use the annotated dataset to create a new sentiment lexicon utilizing corpus-based approach. We use the dataset to train our sentiment model, employing different features and various learning techniques. We also develop a popularity prediction model that predicts the ranking scores for images based on their popularity.

We apply visual, textual, and contextual features in developing our popularity prediction model. The popularity scores are used as indicators for image relevance to the event when determining which images will be included in the event summarization. The list of tweets and images that belong to the event are ranked and a sub-set is selected to represent the subjective multimedia summarization. Extensive experiments are conducted to evaluate the effectiveness of each model and component introduced in our work. We evaluate the image popularity prediction model using different setting of experiments as usually seen on social media. The performance of the sentiment classification model is evaluated on binary and multi-class classification settings beside cross-data experiment setting. The sub-event detection method, image type classifier, and the generated summary are all evaluated using real world data. The results of the conducted experiments show the effectiveness of the developed models throughout this thesis.

The objective of the research conducted throughout this thesis is to summarize the reactions of public as the event evolves over time. The subjective summarization provides an insight into how people react to an event based on their mindset and point of views. Thus, the same event can be described in several different ways, depending on individual experiences, expectations, and perspectives. Our work clearly shows that how football fans react to events is based on their preferred team, and how their feelings and mindsets are directly affected by what happens as the match develops over time, impacting their reactions. Our approach is not limited to football-based events, and could be generalized to include various types of sporting events and other sorts of planned events. The adaptation for other event types would require slight modifications of the length of the time window that is used to monitor the changes in social stream activities based on whether they are long-term or short-term events. Tracking and summarizing human sentiment and reactions during events could be tooled by authorities for riots preventions during and after sport games, understanding public opinions towards political issues, and understanding aspects that affect users' opinions and sentiment, and in turn, their behavior over the course of an event.

7.2 Possible Future Directions

Improving our proposed approach of subjective multimedia event-based summarization could be made considering different components. In terms of the sentiment analysis model, we explored the context-based word embedding features, yet adopting sentiment-based features could lead to better sentiment classifier performance and be a practical, next-step

progression of our work. We also plan to integrate domain expert knowledge to improve our football sentiment lexicon. In addition, a larger dataset is needed for images classification based on synthetic and natural classes. Enriching the image dataset with diverse image types will contribute in providing more consistent classifiers that will benefit a wide range of applications.

Our current work focuses on analyzing the sentiment expressed in fan messages. Further extension to our proposed approach would consider emotion analysis during real world events. This version of our work is built for uni-language, English. Integrating other languages for sentiment analysis and summarization will enrich our work, expanding our resources, and provide access to many more perspectives in regards to an event. Also, we plan to extend our approach to automatically detect events on social media using event detection mechanism in order to make our work suitable for summarizing unplanned events as well as planned events. Additionally, linking event-based data from diverse social media platforms and compiling them into one summary would generate an extensive offline multimedia event-based summarization.

References

[1] Firoj Alam, Ferda Ofli, and Muhammad Imran. Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of hurricanes harvey, irma, and maria. *Behaviour and Information Technology*, 0(0):1–31, 2019.

[2] 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, April 2018.

[3] Luiz Oliveira Alves, Matheus Maciel, Lesandro Ponciano, and Andrey Brito. Assessing the impact of the social network on marking photos as favorites in flickr. In *Proceedings of the 18th Brazilian Symposium on Multimedia and the Web*, WebMedia '12, pages 79–82, 2012.

[4] Flora Amato, Aniello Castiglione, Fabio Mercorio, Mario Mezzanzanica, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperl. Multimedia story creation on social networks. *Future Generation Computer Systems*, 86:412 – 420, 2018.

[5] Javed A. Aslam and Mark Montague. Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 276–284, 2001.

[6] Ben Athiwaratkun, Andrew Wilson, and Anima Anandkumar. Probabilistic FastText for multi-sense word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1–11, July 2018.

[7] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, 2010.

[8] Alexandra Balahur. Sentiment analysis in social media texts. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 120–128, 2013.

[9] Peiman Barnaghi, Parsa Ghaffari, and John G Breslin. Text analysis and sentiment polarity on fifa world cup 2014 tweets. In *Conference ACM SIGKDD*, volume 15, pages 10–13, 2015.

[10] Peiman Barnaghi, Parsa Ghaffari, and John G Breslin. Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. In *IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 52–57, 2016.

[11] Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, August 2017.

[12] Subhabrata Bhattacharya, Behnaz Nojavanasghari, Tao Chen, Dong Liu, Shih-Fu Chang, and Mubarak Shah. Towards a comprehensive computational model for aesthetic assessment of videos. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 361–364. ACM, 2013.

[13] Jingwen Bian, Yang Yang, Hanwang Zhang, and Tat-Seng Chua. Multimedia summarization for social events in microblog stream. *IEEE Transactions on Multimedia*, 17(2):216–228, Feb 2015.

[14] Benjamin Bischke, Damian Borth, and Andreas Dengel. Large-scale social multimedia analysis. In *Big Data Analytics for Large-Scale Multimedia Search*, pages 157–178. Wiley, 2019.

[15] Roi Blanco, Harry Halpin, Daniel M. Herzig, Peter Mika, Jeffrey Pound, Henry S. Thompson, and Thanh Tran Duc. Repeatable and reliable search system evaluation using crowdsourcing. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 923–932, 2011.

[16] 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.

[17] 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.

[18] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001.

[19] Juergen Bross and Heiko Ehrig. Automatic construction of domain and aspect specific sentiment lexicons for customer review mining. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, CIKM '13, pages 1077–1086, 2013.

[20] Elia Bruni, Nam Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49(1):1–47, January 2014.

[21] Emma Byrne and David Corney. Sweet fa: Sentiment, swearing and soccer. In *ICMR2014 1st Workshop on Social Multimedia and Storytelling.*, 2014.

[22] Hongyun Cai, Yang Yang, Xuefei Li, and Zi Huang. What are popular: Exploring twitter features for event detection, tracking and visualization. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 89–98, 2015.

[23] Erion Çano and Maurizio Morisio. Word embeddings for sentiment analysis: A comprehensive empirical survey. *arXiv preprint*, 2019.

[24] Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. Adapting ranking svm to document retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, pages 186–193, 2006.

[25] Spencer Cappallo, Thomas Mensink, and Cees G.M. Snoek. Latent factors of visual popularity prediction. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 195–202, 2015.

[26] Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P. Gummadi. Characterizing social cascades in flickr. In *Proceedings of the 1st Workshop on Online Social Networks*, WOSN '08, pages 13–18, 2008.

[27] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of*

the 18th International Conference on World Wide Web, WWW '09, pages 721–730, 2009.

[28] Deepayan Chakrabarti and Kunal Punera. Event summarization using tweets. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[29] Roshni Chakraborty, Maitry Bhavsar, Sourav Kumar Dandapat, and Joydeep Chandra. Tweet summarization of news articles: An objective ordering-based perspective. *IEEE Transactions on Computational Social Systems*, 6(4):761–777, Aug 2019.

[30] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[31] Freddy Chua and Sitaram Asur. Automatic summarization of events from social media. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2013.

[32] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 48:1–48:9, 2009.

[33] David Corney, Carlos Martn Dancausa, and Ayse Goker. Two sides to every story: Subjective event summarization of sports events using twitter. In *ICMR 2014, 1st International Workshop on Social Multimedia and Storytelling (SoMuS 2014), CEUR Proceedings*, 04 2014.

[34] Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics, 2013.

[35] Jan Deriu, Maurice Gonzenbach, Fatih Uzdilli, Aurelien Lucchi, Valeria De Luca, and Martin Jaggi. SwissCheese at SemEval-2016 task 4: Sentiment classification using an ensemble of convolutional neural networks with distant supervision. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1124–1128, 01 2016.

[36] Wenwen Dou, Xiaoyu Wang, William Ribarsky, and Michelle Zhou. Event detection in social media data. In *IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content*, pages 971–980, 2012.

[37] Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. Evaluation of gist descriptors for web-scale image search. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 19:1–19:8, 2009.

[38] Weifu Du, Songbo Tan, Xueqi Cheng, and Xiaochun Yun. Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 111–120, 2010.

[39] Susan Dumais, John Platt, David Heckerman, and Mehran Sahami. Inductive learning algorithms and representations for text categorization. In *Proceedings of the 7th International Conference on Information and Knowledge Management*, pages 148–155, 1998.

[40] Abdulmotaleb El Saddik. Digital twins: The convergence of multimedia technologies. *IEEE MultiMedia*, 25(2):87–92, 2018.

[41] Abdulmotaleb El Saddik, Hawazin Badawi, Roberto Alejandro Martinez Velazquez, Fedwa Laamarti, Rogelio Gámez Diaz, Namrata Bagaria, and Juan Sebastian Arteaga-Falconi. Dtwins: A digital twins ecosystem for health and well-being. *MMTC Communications-Frontiers*, 14(2):39–46, 2019.

[42] André Luiz Firmino Alves, Cláudio de Souza Baptista, Anderson Almeida Firmino, Maxwell Guimarães de Oliveira, and Anselmo Cardoso de Paiva. A comparison of svm versus naive-bayes techniques for sentiment analysis in tweets: A case study with the 2013 fifa confederations cup. In *Proceedings of the 20th Brazilian Symposium on Multimedia and the Web*, WebMedia '14, pages 123–130, 2014.

[43] Francesco Gelli, Tiberio Uricchio, Marco Bertini, Alberto Del Bimbo, and Shih-Fu Chang. Image popularity prediction in social media using sentiment and context features. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 907–910, 2015.

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

[45] Maria Giatsoglou, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch. Chatzisavvas. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214 – 224, 2017.

[46] Mehreen Gillani, Muhammad U. Ilyas, Saad Saleh, Jalal S. Alowibdi, Naif Aljohani, and Fahad S. Alotaibi. Post summarization of microblogs of sporting events. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 59–68, 2017.

[47] Alec Go, Richa Bhayani, and Lei Huang. Sentiment classification using distant supervision. *Technical report, Stanford*, 2009.

[48] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[49] Jonathan Gratch, Gale Lucas, Nikolaos Malandrakis, Evan Szablowski, Eli Fessler, and Jeffrey Nichols. Goaalll!: Using sentiment in the world cup to explore theories of emotion. In *Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 898–903, Sept 2015.

[50] Bin Guo, Yi Ouyang, Cheng Zhang, Jiafan Zhang, Zhiwen Yu, Di Wu, and Yu Wang. Crowdstory: Fine-grained event storyline generation by fusion of multi-modal crowdsourced data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):55:1–55:19, September 2017.

[51] LI Hang. A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94(10):1854–1862, 2011.

[52] Marko Heikkila and Matti Pietikainen. A texture-based method for modeling the background and detecting moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):657–662, April 2006.

[53] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132, 1999.

[54] Liang-Chi Hsieh, Winston H Hsu, and Hao-Chuan Wang. Investigating and predicting social and visual image interestingness on social media by crowdsourcing. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4309–4313, May 2014.

[55] Liang-Chi Hsieh, Ching-Wei Lee, Tzu-Hsuan Chiu, and Winston Hsu. Live semantic sport highlight detection based on analyzing tweets of twitter. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo*, pages 949–954, July 2012.

[56] Chih-Chung Hsu, Li-Wei Kang, Chia-Yen Lee, Jun-Yi Lee, Zhong-Xuan Zhang, and Shao-Min Wu. Popularity prediction of social media based on multi-modal feature mining. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2687–2691, 2019.

[57] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, 2004.

[58] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen. Local binary patterns and its application to facial image analysis: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):765–781, Nov 2011.

[59] David Inouye and Jugal K Kalita. Comparing twitter summarization algorithms for multiple post summaries. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 298–306, Oct 2011.

[60] Said Jai-Andaloussi, Imane El Mourabit, Nabil Madrane, Samia Benabdellah Chaouni, and Abderrahim Sekkaki. Soccer events summarization by using sentiment analysis. In *Proceedings of the 2015 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 398–403, Dec 2015.

[61] Said Jai-Andaloussi, Aboulaite Mohamed, Nabil Madrane, and Abderrahim Sekkaki. Soccer video summarization using video content analysis and social media streams. In *Proceedings of the 2014 IEEE/ACM International Symposium on Big Data Computingg*, pages 1–7, Dec 2014.

[62] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[63] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM 14, pages 675–678, 2014.

[64] Yu-Gang Jiang, Yanran Wang, Rui Feng, Xiangyang Xue, Yingbin Zheng, and Hanfang Yang. Understanding and predicting interestingness of videos. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, AAAI13, pages 1113–1119, 2013.

[65] Yu-Gang Jiang, Baohan Xu, and Xiangyang Xue. Predicting emotions in user-generated videos. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, AAAI14, pages 73–79.

[66] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 133–142, 2002.

[67] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., 2009.

[68] Peipei Kang, Zehang Lin, Shaohua Teng, Guipeng Zhang, Lingni Guo, and Wei Zhang. Catboost-based framework with additional user information for social media popularity prediction. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2677–2681, 2019.

[69] Lyndon S. Kennedy, Shih-Fu Chang, and Igor V. Kozintsev. To search or to label?: Predicting the performance of search-based automatic image classifiers. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, MIR '06, pages 249–258, 2006.

[70] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. What makes an image popular? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 867–876, 2014.

[71] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762, 2014.

[72] 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.

[73] Chahine Koleejan, Hiroya Takamura, and Manabu Okumura. Generating objective summaries of sports matches using social media. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, pages 353–357, 2019.

[74] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg. In *Proceedings of the 5th International AAAI conference on weblogs and social media*, 2011.

[75] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105. 2012.

[76] Mitsumasa Kubo, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. Generating live sports updates from twitter by finding good reporters. In *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 527–534, Nov 2013.

[77] Kevin Labille, Sultan Alfarhood, and Susan Gauch. Estimating sentiment via probability and information theory. In *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, pages 121–129, 2016.

[78] Kevin Labille, Susan Gauch, and Sultan Alfarhood. Creating domain-specific sentiment lexicons via text mining. In *Proceeding of the Workshop Issues Sentiment Discovery Opinion Mining (WISDOM)*, 2017.

[79] Matthew Lease. On quality control and machine learning in crowdsourcing. In *Proceedings of the 11th AAAI Conference on Human Computation*, AAAIWS'11, pages 97–102, 2011.

[80] Michael Lee. *NBA memes: The role of fan image macros within the online NBA fan community*. PhD thesis, Queensland University of Technology, 2017.

[81] Kristina Lerman and Laurie Jones. Social browsing on flickr. In *Proceedings of the International Conference on Weblogs and Social Media*, 2006.

[82] Rainer W. Lienhart and Alexander Hartmann. Classifying images on the web automatically. *Journal of Electronic Imaging*, 11(4):445 – 454, 2002.

[83] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, July 2004.

[84] Marek Lipczak, Michele Trevisiol, and Alejandro Jaimes. Analyzing favorite behavior in flickr. In *Advances in Multimedia Modeling*, pages 535–545. 2013.

[85] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[86] Bingwei Liu, Erik Blasch, Yu Chen, Dan Shen, and Genshe Chen. Scalable sentiment classification for big data analysis using nave bayes classifier. In *Proceedings of the 2013 IEEE International Conference on Big Data*, pages 99–104, Oct 2013.

[87] Ning Liu, Benyu Zhang, Jun Yan, Zheng Chen, Wenyin Liu, Fengshan Bai, and Leefeng Chien. Text representation: from vector to tensor. In *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM'05)*, Nov 2005.

[88] Annie Louis and Todd Newman. Summarization of business-related tweets: A concept-based approach. In *Proceedings of COLING 2012: Posters*, pages 765–774, December 2012.

[89] Nikolaos Malandrakis, Michael Falcone, Colin Vaz, Jesse James Bisogni, Alexandros Potamianos, and Shrikanth Narayanan. SAIL: Sentiment analysis using semantic similarity and contrast features. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 512–516, August 2014.

[90] Gonçalo Marcelino, Ricardo Pinto, and João Magalhães. Ranking news-quality multimedia. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, pages 10–18, 2018.

[91] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 227–236, 2011.

[92] Philip J. McParlane, Andrew James McMinn, and Joemon M. Jose. "picture the scene...";: Visually summarising social media events. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, CIKM '14, pages 1459–1468, 2014.

[93] Philip J. McParlane, Yashar Moshfeghi, and Joemon M. Jose. "nobody comes here anymore, it's too crowded"; predicting image popularity on flickr. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 385–391, 2014.

[94] Polykarpos Meladianos, Giannis Nikolentzos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. Degeneracy-based real-time sub-event detection in twitter

stream. In *Proceedings of the 9th international AAAI conference on web and social media*, 2015.

[95] Polykarpos Meladianos, Christos Xypolopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. An optimization approach for sub-event detection and summarization in twitter. In *Advances in Information Retrieval*, pages 481–493, 2018.

[96] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint*, 2013.

[97] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRC-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 321–327, June 2013.

[98] Aminu Muhammad, Nirmalie Wiratunga, Robert Lothian, and Richard Glassey. Domain-based lexicon enhancement for sentiment analysis. In *Proceedings of the BCS SGAI Workshop on Social Media Analysis*, 2013.

[99] Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, pages 1–18, 2016.

[100] Nic Newman. The rise of social media and its impact on mainstream journalism. Technical report, 2009.

[101] Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, pages 189–198, 2012.

[102] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC Workshop on 'Mak-ing Sense of Microposts': Big things come in smallpackages*, volume abs/1103.2903, pages 93–98, 2011.

[103] Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part IV*, pages 490–503. Springer, 2006.

[104] Stefanie Nowak and Stefan Rüger. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Pro-*

ceedings of the International Conference on Multimedia Information Retrieval, MIR '10, pages 557–566, 2010.

[105] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, Jul 2002.

[106] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, May 2001.

[107] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*. European Languages Resources Association (ELRA), 2010.

[108] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *roceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[109] Swit Phuvipadawat and Tsuyosh Murata. Breaking news detection and tracking in twitter. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 120–123, Aug 2010.

[110] Shengsheng Qian, Tianzhu Zhang, and Changsheng Xu. Multi-modal multi-view topic-opinion mining for social event analysis. In *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, pages 2–11, 2016.

[111] Xueming Qian, Mingdi Li, Yayun Ren, and Shuhui Jiang. Social media based event summarization by user-text-image co-clustering. *Knowledge-Based Systems*, 2018.

[112] M Elena Renda and Umberto Straccia. Web metasearch: rank vs. score based rank aggregation methods. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 841–846, 2003.

[113] Seyed Mahdi Rezaeinia, Rouhollah Rahmani, Ali Ghodsi, and Hadi Veisi. Sentiment analysis based on improved pre-trained word embeddings. *Expert Systems with Applications*, 117:139 – 147, 2019.

[114] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of machine learning research*, 5:101–141, December 2004.

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

[116] Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, August 2014.

[117] Mickael Rouvier and Benoit Favre. SENSEI-LIF at SemEval-2016 task 4: Polarity embedding fusion for robust sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 202–208, June 2016.

[118] Zafar Saeed, Rabeeh Ayaz Abbasi, Onaiza Maqbool, Abida Sadaf, Imran Razzak, Ali Daud, Naif Radi Aljohani, and Guandong Xu. Whats happening around the world? a survey and framework on event detection techniques on twitter. *Journal of Grid Computing*, 17(2):279–312, 2019.

[119] Horacio Saggion and Thierry Poibeau. Automatic text summarization: Past, present and future. In *Multi-source, Multilingual Information Extraction and Summarization*. Springer, 2013.

[120] Jose San Pedro and Stefan Siersdorfer. Ranking and classifying attractiveness of photos in folksonomies. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 771–780, 2009.

[121] Manos Schinas, Symeon Papadopoulos, Yiannis Kompatsiaris, and Pericles A. Mitkas. Mgraph: multimodal event summarization in social media using topic models and graph-based ranking. *International Journal of Multimedia Information Retrieval*, 5(1):51–69, Mar 2016.

[122] Linda G. Shapiro and George Stockman. *Computer Vision*. Prentice Hall PTR, 1st edition, 2001.

[123] Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. Summarizing microblogs automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 685–688, June 2010.

[124] Credell Simeon, Howard J Hamilton, and Robert J Hilderman. Word segmentation algorithms with lexical resources for hashtag classification. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 743–751, 2016.

[125] Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, MULTIMEDIA '05, pages 399–402, 2005.

[126] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[127] 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 1st Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 53–63, 2011.

[128] Dario Stojanovski, Gjorgji Strezoski, Gjorgji Madjarov, and Ivica Dimitrovski. Finki at SemEval-2016 task 4: Deep learning architecture for twitter sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 149–154, June 2016.

[129] Shiliang Sun, Chen Luo, and Junyu Chen. A review of natural language processing techniques for opinion mining systems. *Information Fusion*, 36:10 – 25, 2017.

[130] Gabor Szabo and Bernardo A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, August 2010.

[131] Yuuki Tagawa and Kazutaka Shimada. *Sports Game Summarization Based on Subevents and Game-Changing Phrases*, pages 65–80. Springer International Publishing, 2018.

[132] Anthony Tang and Sebastian Boring. # epicplay: Crowd-sourcing sports video highlights. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1569–1572, 2012.

[133] Duyu Tang, Bing Qin, and Ting Liu. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432, September 2015.

[134] Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. Efficient object category recognition using classemes. In *European Conference on Computer Vision (ECCV)*, pages 776–789, September 2010.

[135] Jian Tu, Zuxuan Wu, Qi Dai, Yu-Gang Jiang, and Xiangyang Xue. Challenge huawei challenge: Fusing multimodal features with deep neural networks for mobile video annotation. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2014.

[136] Masoud Valafar, Reza Rejaie, and Walter Willinger. Beyond friendship graphs: A study of user interactions in flickr. In *Proceedings of the 2nd ACM Workshop on Online Social Networks*, WOSN '09, pages 25–30, 2009.

[137] Chris van der Lee, Emiel Krahmer, and Sander Wubben. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 95–104, September 2017.

[138] Guido Van Oorschot, Marieke Van Erp, and Chris Dijkshoorn. Automatic extraction of soccer game events from twitter. In *Proceedings of Detection, Representation and Exploitation of Events in the Semantic Web (DeRiVE 2012)*, 2012.

[139] Roelof van Zwol. Flickr: Who is looking? In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 184–190, 2007.

[140] Roelof van Zwol, Adam Rae, and Lluis Garcia Pueyo. Prediction of favourite photos using social, visual, and textual signals. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 1015–1018, 2010.

[141] Christopher C. Vogt and Garrison W. Cottrell. Fusion via a linear combination of scores. *Information Retrieval*, 1(3):151–173, October 1999.

[142] Fei Wang and Min-Yen Kan. Npic: Hierarchical synthetic image classification using image search and generic features. In *Proceedings of the 5th International Conference on Image and Video Retrieval*, pages 473–482, 2006.

[143] Jenq-Haur Wang, Ting-Wei Liu, Xiong Luo, and Long Wang. An LSTM approach to short text sentiment classification with word embeddings. In *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING)*, pages 214–223, October 2018.

[144] Xinyue Wang. *Real-time Content Identification for Events and Sub-Events from Microblogs.* PhD thesis, Queen Mary University of London, 2016.

[145] Zhaoxia Wang, Victor Joo Chuan Tong, Pingcheng Ruan, and Fang Li. Lexicon knowledge extraction with sentiment polarity computation. In *Proceedings of the 16th International Conference on Data Mining Workshops (ICDMW)*, pages 978–983, Dec 2016.

[146] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, and Xin Li. An optimal svm-based text classification algorithm. In *Proceedings of the 2006 International Conference on Machine Learning and Cybernetics*, pages 1378–1381, Aug 2006.

[147] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354, 2005.

[148] Bo Wu, Tao Mei, Wen-Huang Cheng, and Yongdong Zhang. Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, AAAI16, pages 272–278. AAAI Press, 2016.

[149] Toshihiko Yamasaki, Shumpei Sano, and Kiyoharu Aizawa. Social popularity score: Predicting numbers of views, comments, and favorites of social photos using only annotations. In *Proceedings of the 1st International Workshop on Internet-Scale Multimedia Management*, WISMM '14, pages 3–8, 2014.

[150] Xiao Yang, Craig Macdonald, and Iadh Ounis. Using word embeddings in twitter election classification. *Information Retrieval Journal*, 21(2):183–207, Jun 2018.

[151] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schtze. Comparative study of cnn and rnn for natural language processing. *arXiv preprint*, 2017.

[152] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, 13:55–75, 08 2018.

[153] Felix Yu, Liangliang Cao, Rogerio Feris, John Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 771–778, June 2013.

[154] Hwanjo Yu and Sungchul Kim. *SVM Tutorial — Classification, Regression and Ranking*, pages 479–506. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[155] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, September 2015.

[156] Siqi Zhao, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. Human as real-time sensors of social and physical events: A case study of twitter and sports games. *arXiv preprint*, 2011.

[157] Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. Learning continuous word embedding with metadata for question retrieval in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 250–259, July 2015.

[158] Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, and Julio Gonzalo. Towards real-time summarization of scheduled events from twitter streams. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*, HT '12, pages 319–320, 2012.