Towards a Tweet Analysis System to study Human Needs during COVID-19 pandemic

Zijian Long

Thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
M.A.Sc. degree in Electrical and Computer Engineering

School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

Abstract

Governments and municipalities need to understand their citizens' psychological needs in critical times and dangerous situations. COVID-19 brings lots of challenges to deal with. We propose NeedFull, an interactive and scalable tweet analysis platform, to help governments and municipalities to understand residents' real psychological needs during those periods. The platform mainly consists of four parts: data collection module, data storage module, data analysis module and data visualization module. The whole process of how data flows in the system is illustrated as follows: Our crawlers in the data collection module gather raw data from a popular social network website Twitter. Then the data is fed into our human need detection model in the data analysis module before stored into the database. When a user enters a query through the user interface, they will get all the related items in the database by the index system of the data storage module and a comprehensive human needs analysis of these items is then presented and depicted in the data visualization module. We employed the proposed platform to investigate the reaction of people in four big regions including New York, Ottawa, Toronto and Montreal to the ongoing worldwide COVID-19 pandemic by collecting tweets posted during this period. The results show that the most pronounced human need in these tweets is relatedness with 51.32%, followed by autonomy with 22.56% and competence with 18.82%. And the percentages of tweets expressing frustration are larger than those of tweets expressing satisfaction for each psychological need in general.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Abdulmotaleb El Saddik for his guidance and support for my study and research, for his patience, motivation, enthusiasm, and immense knowledge. I am proud to be one of his students and I could not have finished this thesis without him.

I would like to thank the Faculty of Graduate and Postdoctoral Studies, the University of Ottawa for giving me this platform and opportunity.

My sincere thanks also goes to all the members in MCR lab for their kindness and help especially during my dark days. I am gratefully indebted to Rajwa Alharthy, Samah Aloufi and Fatimah Alzamzami for their very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my parents and to my grandparents for providing me with unfailing support and continuous encouragement throughout my years of study.

Contents

1 Introduction **1**

1.1 Background and Motivation 1
1.2 Thesis Objectives 4
1.3 Thesis Statement 5
1.4 Scholarly Output 5
1.5 Thesis Outline 5

2 Related Work **6**

2.1 Sentiment Analysis 6
2.1.1 Document-level Sentiment Analysis 6
2.1.2 Sentence-level Sentiment Analysis 7
2.1.3 Aspect-level Sentiment Analysis 13
2.2 Emotion classification and Human Need Detection 14
2.2.1 Dictionary-based Methods 14
2.2.2 Rule-based Methods 15
2.2.3 Machine Learning-based Methods 16
2.2.4 Multi-label methods 16
2.2.5 Human Need Detection 17
2.3 Human Affective State Analysis Platforms 17
2.4 Affect-aware City 18

3 System Design and Implementation **21**

3.1 Architecture Design 21
3.2 Data Collection 23
3.2.1 Recent tweets collector 24
3.2.2 Historical tweets collector 26
3.3 Data Analysis 28

3.3.1 Data Preprocessing . . . 29
3.3.2 Feature Extraction . . . 29
3.3.3 Normalization . . . 32
3.3.4 Dimensionality Reduction . . . 32
3.3.5 Evaluation of the Models . . . 32
3.4 Data Storage . . . 33
3.5 Data Visualization . . . 35
3.6 System Implementation . . . 37
3.7 Security Assurance System . . . 39
3.7.1 Operating system security . . . 39
3.7.2 Data security . . . 39
3.7.3 Network security audit . . . 39

4 Human needs analysis of COVID-19 pandemic 40
4.1 Human needs analysis of COVID-19 pandemic in New York State . . . 40
4.1.1 Human needs analysis of COVID-19 pandemic in New York State in week 1 . . . 41
4.1.2 Human needs analysis of COVID-19 pandemic in New York State in week 2 . . . 43
4.1.3 Human needs analysis of COVID-19 pandemic in New York State in week 3 . . . 44
4.1.4 Human needs analysis of COVID-19 pandemic in New York State in week 4 . . . 45
4.1.5 Human needs analysis of COVID-19 pandemic in New York State in week 5 . . . 46
4.1.6 Overall human needs analysis of COVID-19 pandemic in New York State . . . 47
4.2 Human needs analysis of COVID-19 pandemic in Ottawa, Toronto and Montreal 51

5 Conclusion and Future Work 53
5.1 Conclusion . . . 53
5.2 Future Work . . . 54

A Glossary of Terms 56

References 58

List of Tables

3.1 Accuracy, Recall (R), Precision (P) and F_{score} for the human need models [83] . 33
3.2 Hardware development environment 38
3.3 Software development environment 39

4.1 Top 10 most frequently used hashtags in tweets during COVID-19 pandemic in New York State using NeedFull platform 42

List of Figures

1.1 Social Media Webs and Apps 2

2.1 Model architecture of FastText for a post with N n-gram features x_1, ..., x_N which are embedded and averaged to form the hidden variable. 9
2.2 TextCNN architecture for an example sentence 10
2.3 Architecture of a typical RNN 11
2.4 Simplified model to explain attention mechanism 12
2.5 Architecture of the proposed system to detect and visualize emotions from geo-tagged Twitter messages [74]. 19

3.1 Architecture of NeedFull platform 22
3.2 Applying for access to Twitter API 24
3.3 Twitter's Search API 25
3.4 Architecture of NeedFull crawler for historical data 27
3.5 An example of DOM 27
3.6 Offline training and testing phase of the human need detection framework[78] . 28
3.7 Design of database in NeedFull platform 35
3.8 Use Case Diagram in UML 36
3.9 User interface of NeedFull: Register page (a), Login page (b), Search page (c), and Result page (d) 37
3.10 MVC Architectural Pattern of Django 38

4.1 Measuring weekly change of satisfaction level during COVID-19 pandemic in New York State using NeedFull platform 41
4.2 Measuring weekly change of frustration level during COVID-19 pandemic in New York State using NeedFull platform 41

4.3 Word cloud generated from tweets during COVID-19 pandemic in New York State using NeedFull platform: WEEK 1 (a), WEEK 2 (b), WEEK 3 (c), WEEK 4 (d) and WEEK 5 (e) 44
4.4 Recognizing life aspects distribution during COVID-19 pandemic in New York State using NeedFull platform 48
4.5 Recognizing human needs distribution during COVID-19 pandemic in New York State using NeedFull platform 48
4.6 Identifying social context types during COVID-19 pandemic in New York State using NeedFull platform 50
4.7 Recognizing human needs distribution during COVID-19 pandemic in Ottawa using NeedFull platform 51
4.8 Recognizing human needs distribution during COVID-19 pandemic in Toronto using NeedFull platform 51
4.9 Recognizing human needs distribution during COVID-19 pandemic in Montreal using NeedFull platform 52

Chapter 1

Introduction

1.1 Background and Motivation

Social media, as one of the best ways for people to share their lives and connect with each other, has already become a significant part of people's lives in 2020 [1]. On average, we spend 144 minutes every day on social media platforms such as Facebook, Twitter, LinkedIn and Pinterest (Figure 1.1) [1]. The rapid development of social media has brought a huge number of users using these social media websites and Apps [2]. For example, there are 330 million monthly active users (MAU) on Twitter [2] in 2020 which is larger than 6 times as many as that in 2010. Moreover, they post around 500 million posts about a vast range of events happening all over the world on Twitter per day. The huge number of users and millions of posts on social media reflect the current changes in the way people acquire and perceive information. Traditional media, such as newspaper, radio, television, which was the main way of how we got information, only provides a one-way communication process. Sometimes it is forced on you when you do not want to see some ads and there is not a community for people to talk about the information. Different from traditional media [3], social media platforms allow users to discover the content of interest, and to convey their opinions, attitudes and emotions to the outside world. For example, they could express their preferences for celebrities and politicians, personal opinions after reading the latest news from various organizations, likes or dislikes of a certain corporate brand and evaluation of some products [4]. This brings the rise of social big data which can be defined as follows [5]: "Those processes and methods that are designed to provide sensitive and relevant knowledge to any user or company from social media data

[1] https://www.broadbandsearch.net/

[2] https://about.twitter.com/company/

sources when data sources can be characterized by their different formats and contents, their very large size, and the online or streamed generation of information."

Figure 1.1: Social Media Webs and Apps

Compared with traditional media, social big data has the following new features [6]. First of all, the limitation of the number of words is one of the most significant features of social big data. Under normal circumstances, the number of words in social big data is limited to no more than 140 characters. This feature requires users to express their emotions or event information they want to share in a concise way [7], [8]. Secondly, unlike the formal texts used by traditional media, the grammar of social media texts is usually informal. There is no restriction on grammar structure and the words they use are more colloquial. When editing the content for social media, users do not need to consider whether the grammatical structure of the sentence is standard, and they can choose any words to express their views based on their preferences. Even sometimes there is only one sentence, one word or even one symbol in their posts. Thirdly, there is massive repetitive and useless information on social media. These noise posts are not only meaningless, but also seriously affect the quality of social media analysis. Fourthly, texts on social media usually appear in the form of a dialogue. After the first post of a related topic is posted, many follow-up posts will be generated in the form of replies, where a large amount of text will be generated, forming a rich contextual background [9]. These contexts are an important basis for understanding and analyzing their posts. In addition, many users prefer to use abbreviations on social media platforms. For example, "LOL" stands for laugh out loud and "ASAP" represents as soon as possible.

Since social media is becoming a more and more important part of people's daily life, ex-

ploring and mining social big data to understand people's various affective states has enormous potential values. First of all, it has social value. Online public opinion refers to the opinions, attitudes and emotions supported by the majority of the people in response to various social problems and phenomena in the Internet environment [10].It is the projection of real public opinion on the Internet. If governments can correctly notice the trend of online public opinion, they will be able to realize the sentiments and needs of the people behind the online public opinion and then provide a proper environment to increase people needs satisfaction level. But on the contrary, if governments cannot control online public opinion in its infancy, subjective opinions and emotions will continue to spread through the Internet, and then causes unpredictable losses [11]. Therefore, it is very important to be able to reasonably guide the direction of public opinion, monitor and restrict people's behavior online. Moreover, it has business value, too. The emergence of social media has greatly changed the mode of people's communication and information transmission, and also made enterprises face more opportunities and challenges. Enterprises can run official accounts on the Internet to gain customers' attention and improve their popularity. During the interaction between enterprises and customers on social media platforms, enterprises can analyze consumers' emotional tendencies and personal needs in a timely manner and obtain improvement suggestions for their products from customers [12]. Moreover, enterprises can also deal with crisis faster by using these platforms.

In order to acquire a thorough analysis of human affective states, the first layer is to identify the sentiment type (positive, neutral and negative). The research of sentiment analysis on social media has been developed for more than ten years and achieved good results. In 2009, Alec Go et al., for the first time, used machine learning methods to automatically conduct sentiment analysis on the text on Twitter through remote supervision, and added emoticons to the system, which greatly improved the accuracy of the system [13]. In 2010, O'Connor et al. counted the words used by consumers on product views and opinion polls between 2008 and 2009, and found that there was a high correlation between the commonly used emotional words for two environments [14]. Pak et al. collected tweets as a corpus to perform sentiment analysis and opinion mining by training classifiers [15]. Bifet et al. analyzed the difficulties of data mining on social data and set a sliding window to solve the problem of imbalance in data [16]. Davidov et al. used the tags and emoticons on Twitter as labels, and built a sentiment classification system through supervised learning. This method did not require too many manual annotations [17]. Based on a probabilistic model, Asli et al. first classified the text on twitter into polar (tweets with positive or negative sentiment) and non-polar (tweets without sentiments), and then used emotional vocabulary to classify the emotional polarity of polar tweets [18].

Textual emotion detection, whose task is to classify a text into one or more predefined emotion categories by extracting the emotional elements in the textual content, has gained a lot of attention in recent years. Generally speaking, the emotional classification methods mainly include dictionary-based methods, rule-based methods, machine learning-based methods, composite methods, and multi-label methods [19].Bollen et al. used a psychometric instrument to extract six emotional states (tension, depression, anger, vigor, fatigue, confusion) from texts, and calculated the value of the six-dimensional emotional vector according to the daily timeline. They compared the calculation results with events that occurred in other media and found out that social events in all life aspects have a significant impact on the public emotion [20]. Moreover, some researches have started to explore the methods to analyze human psychological needs which can recognize the psychological need type (relatedness, competence and autonomy) according to the Human Needs Theories (HNT) [21] and discover if this need is satisfied or not. The awareness of human needs is a significant step to improve people's well-being. In [22], the authors proposed a multi-layered psychological-based reference model to assess citizen needs during any event at any time.

1.2 Thesis Objectives

The goal of this thesis is to develop a human need analysis platform based on Twitter which we refer to as NeedFull. The objectives of developing a well-functioning data analysis platform are shown as follows:

1. We need to develop and implement a well-performed human need detection framework which can be used to automatically analyze downloaded tweets and label these tweets for each model of the framework.

2. Since NeedFull is a social data analysis platform, the modules used to deal with a large number of data need to be built including data collection module, data storage module and data analysis module.

3. To ensure the normal operations of the system and to provide users with a good experience during the operation of the system, some non-functional requirements need to be satisfied, such as user-friendly interfaces and interactive dashboards.

1.3 Thesis Statement

In this work, We propose NeedFull, an interactive and scalable tweet analysis platform, to help governments and municipalities to understand residents' real psychological needs during those periods. The platform mainly consists of four parts: data collection module, data storage module, data analysis module and data visualization module. The four parts interact with each other and provide users with a thorough human needs analysis based on their queries. We employed the proposed platform to investigate the reaction of people in New York State and three big cities in Canada including Ottawa, Toronto and Montreal to the ongoing worldwide COVID-19 pandemic. The main contributions of this work are shown as follows:

- Design and development of a tweet analysis platform NeedFull
- Implementation of a human need detection framework
- Design and development of a crawler which can collect recent and historical data on Twitter
- Design and development of an indexing system of our database

1.4 Scholarly Output

As a result of this work, a journal called "NeedFull – a Tweet Analysis Platform to Study Human Needs During the COVID-19 Pandemic in New York State" has been accepted by IEEE Access.

1.5 Thesis Outline

The remainder of the paper is organized as follows:

- Chapter 2 presents a review of related work of detecting human affective states.
- Chapter 3 describes requirements analysis of our system
- Chapter 4 illustrates system design and implementation.
- Chapter 5 investigates the public reaction to the ongoing worldwide COVID-19 pandemic using our system.
- Chapter 6 concludes our work and discusses future work.

Chapter 2

Related Work

Generally speaking, analyzing human affective states involves three steps: sentiment analysis, emotion detection and human need detection. In this chapter, we will first introduce related techniques for each step and then we will describe some social media analysis platforms using these techniques. In the end, some researches about affect-aware city where public affections can be detected will be presented [23].

2.1 Sentiment Analysis

The first step is sentiment analysis, also known as opinion extraction, is the computational treatment of opinions, sentiments and subjectivity of text are employed [24]. Sentiment analysis techniques can be divided into three categories based on different levels:

2.1.1 Document-level Sentiment Analysis

Document-level sentiment analysis aims to automate the task of classifying a textual review, which is given on a single topic, as expressing a positive or negative sentiment [25]. These approaches are usually categorized as supervised machine learning approaches or unsupervised machine learning approaches. For supervised machine learning approaches, Yan Zhao et al. in 2014 performed experiments about sentiment based on news comments in Chinese including four feature selection methods: Document frequency (DF), Information Gain (IG), Mutual Information (MI) and Chi-square Statistic (CHI); three feature representations: Presence, Term Frequency (TF), Term Frequency weighted by Inverse Document Frequency (TF-IDF) and five learning methods: Naive Bayes (NB), Maximum Entropy (ME), Winnow, C4.5, Support Vector

Machine (SVM) [26]. Different from Yan Zhao's work only focusing on traditional machine learning methods, Lina L et al. combined NB and Neutral Network in their proposed system to analyze sentiment of movie review [27]. Mathews et al. proposed a lexicon based method to perform polarity calculation on the multilingual dataset which consists of a mix of reviews in English and Malayalam for sentiment [28]. The proposed methodology treats both types of lexicons differently and it gives more accurate results for sentiment analysis. In order to solve the problem that the existing researches cannot exploit the deep semantic information of documents, Liu et al. proposed a novel hierarchical neural network model based on dynamic word embeddings (HieNN-DWE) for document level sentiment classification [29]. The model consists of two layers: the first one uses bidirectional gated recurrent unit (BiGRU) and attention mechanism to encode sentences and in the second layer, both BiGRU and convolutional neural network (CNN) are employed to capture features in the sentences. This model outperforms existing methods on four public datasets for document level sentiment classification.

Regarding unsupervised machine learning approaches, an algorithm based on semantic orientation (SO) proposed by Peter D. Turney is applied to classification of reviews about automobiles, banks, movies, and travel destinations in 2002 [30]. The basic idea is that the sentiment of the whole document is determined by SO of words containing sentiment which are already identified according to whether they are closer to the word 'Excellent' or the word 'Poor'. To improve the classification result by the above algorithm, the Semantic Orientation CALculator (SO-CAL), a lexicon based method is proposed by M. Taboada et al. in 2011 where they use different dictionaries for adjectives, nouns, verbs, adverbs, intensification, negation, Irrealis Blocking and text-level features to calculate the value of SO [31]. Compared to previous systems, SO-CAL has some outstanding features that it assigns different weights to the sentences and it allows cut-offs which means the classification results can be divided to n categories.

2.1.2 Sentence-level Sentiment Analysis

Sentence level sentiment analysis is a sub-class analysis of the documents which aims to classify only one sentiment as positive, neutral or negative. Methods of sentence level sentiment analysis mainly include traditional machine learning methods and deep learning methods.

With the development of machine learning, many classic methods for solving large-scale text classification problems have been created gradually [32]. The main process includes two parts: feature engineering and classifiers training. Feature engineering plays an extremely significant role in the traditional sentiment process including text preprocessing, feature extraction, and text representation [33]. As for classifiers, they are mainly designed using some sta-

tistical methods, such as Naive Bayes, Random Forest (RF) and Support Vector Machine. The biggest problem of the traditional sentiment analysis methods is the sparse and high-dimension text representation and manual feature engineering with high cost. The key to applying deep learning methods to sentiment analysis is choosing an appropriate way to represent texts and then taking advantage of the neural network's ability to automatically learn the features of texts to get rid of the troublesome feature engineering.

With respect to words representation, there are two main ways for now: one-hot presentation and distributed representation. One-hot representation is a common, easy way where the word is represented by a long vector whose dimension is the length of all the words with most of the elements being 0 and only one 1 to identify the word. This type of representation is able to solve most of the Natural Language Processing (NLP) tasks incorporated with the traditional sentiment analysis methods [34]. However, the limitations of one-hot representation should not be ignored [35]. First, the dimension of vector becomes higher as the number of words gets bigger which could lead to the curse of dimensionality. Second, any two words are independent which means it is impossible to express relevant information between words at the semantic level. In order to merge the semantic meaning of words into their representation, distributed representation is generated based on the distributional hypothesis by Harris in 1954 which is words with similar context have similar meanings [36]. Word embedding is one of distributed representation methods built on neural networks whose core is modeling the relationship between the target word and its context. Actually, there is another concept called statistical language model which can be used for the similar purpose. The statistical language model takes a sequence of words as a random event and calculates the probability of this sentence by which the information of word order which can be seen as relationship between the target word and its context is kept. So word embedding can be considered as a by-product of training statistical language model at the first place.

In 2013, Tomas Mikolov et al. proposed two novel models: Continuous Bag-of-Words Model (CBOW) and Skip-gram Model shown in fig6 with a simple and practical tool word2vec to provide high-quality representations of words [37]. Based on this, the representation of texts is transformed into a type of continuous dense data like image and speech through word embedding. Since the deep learning algorithms are good at dealing with this kind of data, some algorithms such as CNN that performs well in the image field can also be applied to sentiment analysis. In the following part, I will introduce several deep learning networks used for sentence-level sentiment analysis.

FastText

In 2016, Tomas Mikolov et al. proposed an easy, fast model for text classification called FastText [38]. The model architecture is shown in Figrue 2.1 which only includes three layers: an input layer, a hidden layer and an output layer. The principle of this model is to average all the word vectors of a sentence in the hidden layer which directly connects to the classification layer to get the label for the sentence. In addition, they added some features acquired from n-gram model which can divide a sentence into sequences of N words. Due to the limitation of the structure, FastText can only capture some linear features of the sentences. However, it is still an efficient model for cursory large-scale text classification.

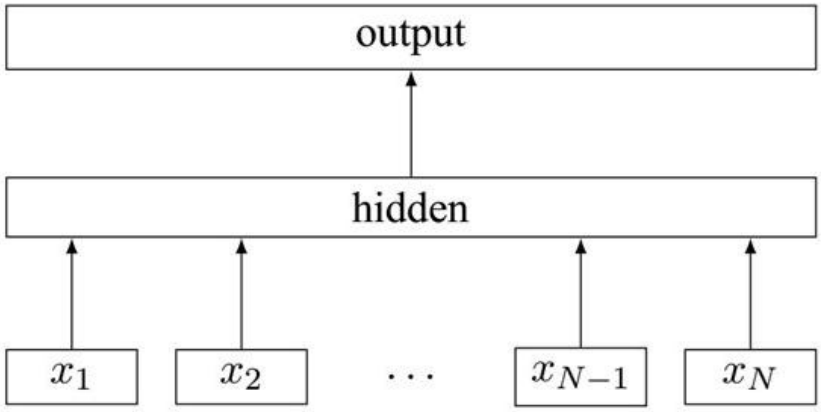


Figure 2.1: Model architecture of FastText for a post with N n-gram features x_1, ..., x_N which are embedded and averaged to form the hidden variable.

TextCNN

The order of words in a sentence plays an important role in understanding the sentiment of a sentence. To capture more information of the order of words in a sentence, TextCNN was proposed by Yoon Kim in 2014 [39]. He applied word embedding on an example sentence with 9 words and ended up with a 9 by 6 matrix to present the sentence. The structure of TextCNN is shown in Figrue 2.2. The input layer connects to a one-dimensional convolutional layer with 6 filters: 2 for each region size (2, 3, 4). After that comes a 1-max pooling layer so that sentences with different lengths can have the same length after passing through the pooling layer. A fully connected softmax layer is added in the end of the structure to output the probability of each category. To better understand the above process, some important concepts

will be explained. The first one is features of words which is represented by word embedding. There are two ways that can be applied here: the static way and the non-static way [40]. The static method usually uses pre-trained word2vec to present words and the training process does not update the word vectors [41]. Therefore, it is essentially a type of transfer learning which is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task [42]. Especially when the quantity of data is relatively small, the static word vector tends to perform well. The non-static method updates the word vector during training the model. It initializes the word vector with a pre-trained word2vec vector like the static method and then keeps adjusting the word vector during training which can accelerate convergence. If there is sufficient training data, direct random initialization of the word vector is also feasible. As for channels, the input channels of text are usually different ways of embedding compared to the RGB (red, green, blue) channels used in images which stands for red, green and blue channel respectively. In practice, static word vectors and non-static fine-tunning word vectors are also used as different channels. Pooling layers are usually employed in CNN to retain the main features and reduce parameters in order to prevent over-fitting [43]. Nal Kalchbrenner et al. proposed dynamic k-max pooling where k pieces of important information are kept [44]. For example, here is a sentence, 'I love the environment here, but it cost me too much money'. Although the first half reflects the positive emotion, the whole text expresses the negative emotion. k-max pooling can capture this type of information well.

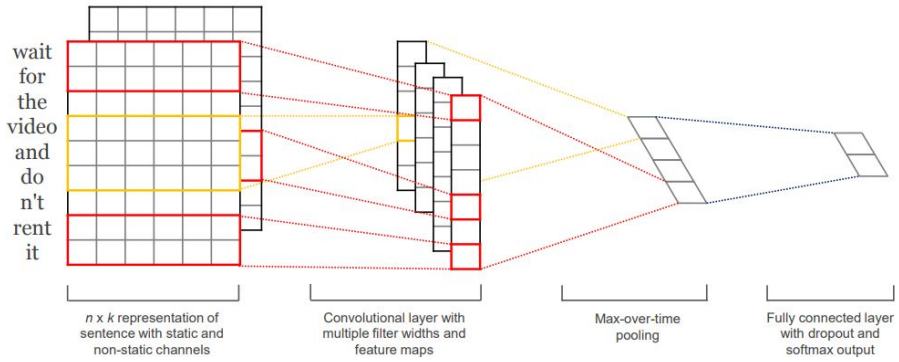


Figure 2.2: TextCNN architecture for an example sentence

TextRNN

Although TextCNN can perform well in many NLP tasks, there are still some problems in it. First, TextCNN cannot model long sequence information due to the nature of CNN. Second, the adjustment of hyperparameters such as the size of filters is cumbersome. Therefore, Recurrent Neural Network (RNN), which is good at dealing with sequence data like texts, has got much attention from researchers [45]. What a typical RNN looks like is shown in Figure 2.3. It shows a RNN being unfolded into a full network. By unfolded we simply mean that we write out the network for the complete sequence. For example, if the sequence we care about is a sentence of 5 words, the network would be unrolled into a 5-layer neural network, one layer for each word. The formulas that govern the computation happening in a RNN are as follows: x_t is the input at time step t. For example, x_1 could be a one-hot vector corresponding to the second word of a sentence. s_t is the hidden state at time step t. It's the "memory" of the network. s_t is calculated based on the previous hidden state and the input at the current step. s_{-1} is required to calculate the first hidden state, is typically initialized to all zeroes. o_t is the output at step t. If we want to predict the next word in a sentence, it would be a vector of probabilities across our vocabulary.

Moreover, Long et al. proposed a RNN-based model for text classification where Long short-term memory network (LSTM) is used to solve this issue of learning long-term dependencies [46]. For now, basic RNN and other versions of RNN like LSTM, Gated Recurrent Unit (GRU) are widely used in the field of NLP such as sequence annotation, Name Entity Recognition (NER) and sequence to sequence models [47].

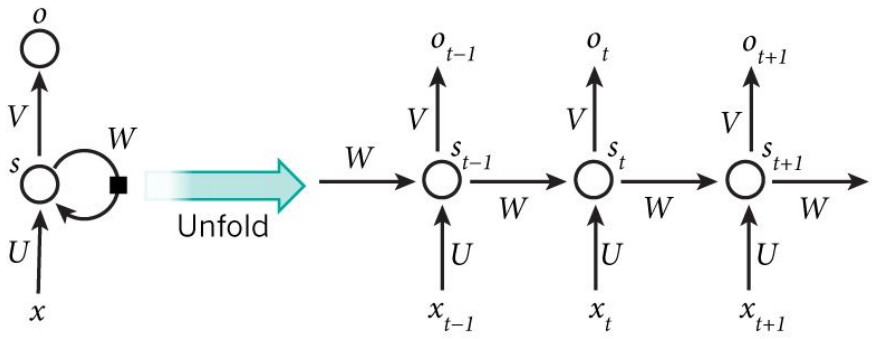


Figure 2.3: Architecture of a typical RNN

TextRNN + attention mechanism

The above CNN and RNN models have a deficiency that the outcome is not intuitive which means people do not understand how the networks get the result. Due to this end-to-end process, the interpretability of the results is not good, especially when analyzing bad cases.

In 2014, the attention mechanism is first applied to images by Volodymyr et al. [48]. The motivation is inspired by the human's attention mechanism. When people observe images, they do not look at each pixel of the entire image at a time. Most of them focus on specific parts of the image according to their needs. Moreover, humans are able to learn from the previously observed images to predict which part of the image more attention should be paid to. The attention mechanism can also be applied to NLP. For example, we have the sentence "How was your day", which we would like to translate to the French version "Comment se passe ta journée". What the Attention component of the network will do for each word in the output sentence is map the important and relevant words from the input sentence and assign higher weights to these words, enhancing the accuracy of the output prediction shown in Figure 2.4. Moreover, Dzmitry et al. added the attention mechanism to Neural Machine Translation (NMT) which is a typical seq2seq model [49]. The traditional NMT uses two RNNs: One is used to encode the source sentence into a fixed-dimensional vector and the other one is to decode the vector to the target language. They used a perceptron to associate each word of the target sentence with each word of the source sentence to get a probability distribution of attention. By doing this, the accuracy of the classification results are improved markedly.

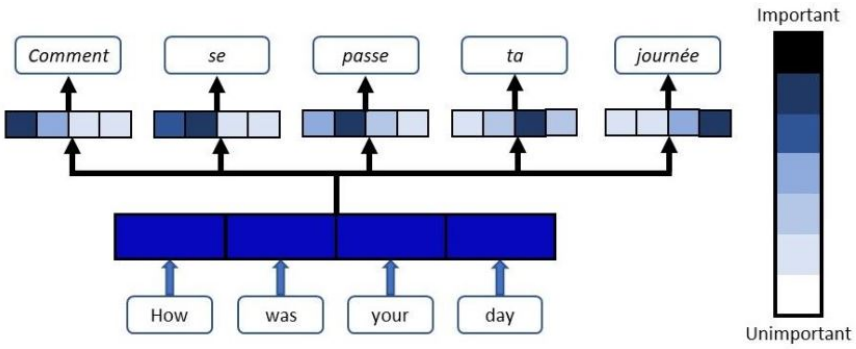


Figure 2.4: Simplified model to explain attention mechanism

TextRCNN

RNN and CNN, as the basic neural networks for text classification, have their own advantages and limitations. RNN is good at processing sequence data and can take into account the context

information of the sentence. However, RNN is a biased model where the words that appear in the late part of a sentence are more important due to the structure of RNN. This may affect the final classification result as important words may be anywhere in a sentence. CNN is an unbiased model, which can obtain the most important features through maximum pooling. However, the sliding window size of CNN is not easy to determine. If it is too small or too large, it will cause important information to be lost. In order to solve the limitations of RNN and CNN, Siwei et al. proposed a new network architecture, which uses a bidirectional loop structure to obtain context information [50]. This model can reduce more noise than traditional window-based neural networks, and can preserve the order of words in sentences. Moreover, the maximum pooling layer is used to automatically determine which parts are more important in the sentiment classification process.

2.1.3 Aspect-level Sentiment Analysis

The research on document-level and sentence-level sentiment analysis is relatively early, but the results of the analysis are not ideal when dealing with complex sentences where multiple aspects show up which is quite common in the real world. Especially when it comes to reviews of products, the overall sentiment provide by document-level or sentence level analysis does not match the attitude towards each aspect of the products. For example, here is a sentence, 'The color is brilliant but the quality is not good' the sentiment of which is positive regarding color but it is negative as for quality. Obviously, only one sentiment is not enough for analyzing two aspects of the product.

Aspect level sentiment analysis, also known as aspect-based sentiment analysis (ABSA), is a text analysis technique that breaks down text into aspects (attributes or components of a product or service), and then allocates each one a sentiment type (positive, negative or neutral) [51]. The basic steps of aspect level sentiment analysis include identification, classification, and aggregation. Concretely, identification is used to find the aspects involved in the text and classification is employed for classifying the sentiment of each aspect as positive, neutral and negative based on the methods from sentence-level analysis. At the end, a general analysis is provided by aggregating the sentimental values. Cruz et al. proposed a domain-oriented approach where they mapped emotion words of different domains into corresponding lexicons to figure out the way people express their sentiments about various domains[52]. In [53], the authors proposed an aspect-level sentiment analysis method based on ontologies in the diabetes domain.

2.2 Emotion classification and Human Need Detection

Emotion classification, whose task is to classify a text into one or more predefined emotion categories by extracting the emotional elements in the textual content, has gained a lot of attention in recent years. Generally speaking, the emotion classification methods mainly include dictionary-based methods, rule-based methods, machine learning-based methods, and multi-label methods [19].

2.2.1 Dictionary-based Methods

The dictionary-based methods mainly use emotional dictionaries to extract emotional expression keywords in texts, and classify texts based on those words. In 2005, Ma et al. proposed a dictionary-based emotion classification method which was applied to an instant messaging system [54]. The method used keywords to identify emotion-related content first, and then detected the emotional meaning using syntactic features. After that, these generated emotions were used to adjust the speech synthesis function to help users better communicate with other remote users. Based on this method, Aman et al. proposed a better emotion classification method employing emotional intensity annotation, which assigned different intensities to emotional words [55]. This method has reached an accuracy rate of more than 66% in the emotional classification task of blog corpus.

Because emotional words in emotional dictionaries are usually related to certain domain, time and language, the same words may express completely different emotions in different domains, times and languages. However, previous methods did not consider these factors when constructing emotional dictionaries. Therefore, the results were not ideal in the cross-domain, cross-time, cross-language textual emotion classification tasks.

To solve domain dependence, Yang et al. proposed a classification method based on specific domain sentiment dictionaries [56]. This method used emotion-aware Latent Dirichlet Allocation (EaLDA) model to construct different emotional dictionaries for specific fields with predefined emotions. The EaLDA model utilized a set of domain-independent minimal seed words as a prior knowledge to construct specific domain dictionaries. This method could effectively improve the results for textual emotion classification tasks in SemEval-2007 dataset where it reached an accuracy rate of 36.85% for detecting sadness. To explore the impact of time dependence on emotion classification, Golder et al. employed emotional dictionaries to collect tweets published by users from different regions and cultures around the world, and analyzed the emotions expressed in millions of tweets [57]. The results clearly showed that people's emotions presented periodic patterns in different time zones. With regard to language

dependence, the most common method is to build a specific emotional dictionary for certain language. For some languages which lack corpus resources, it is difficult to obtain corpus material used to construct the emotional dictionary. Xu et al. proposed WordNet-Affect, a Chinese emotional dictionary based on the English emotion dictionary [58]. The method first translated all English words in the English sentiment dictionary into Chinese, then built a bilingual undirected graph for each emotional category by means of a Chinese synonym dictionary called "Synonym Word Forest" and excluded non-emotional words to get a set of emotional seed words. Finally, the dictionary was expanded by adding synonyms of emotional words.

In general, the dictionary-based methods can reflect the unstructured features of textual data. In the case of high coverage of the emotional words and high accuracy of the annotation in the dictionary, the classification results are ideal [59]. However, these methods depend too much on domains, times and languages and it is difficult to capture new generated words. Therefore how to construct a high-quality emotional dictionary is a challenge for researchers.

2.2.2 Rule-based Methods

In rule-based methods, classification rules need to be defined in advance, including expert annotation, professional dictionaries, statistical methods, etc. In 2008, Strapparava et al. proposed an emotion classification method based on semantic meaning [60]. They employed Latent Semantic Analysis (LSA) to calculate the semantic similarity between general words and emotional words, and then classified news headlines according to semantic similarity. This method was used in emotional classification tasks in news corpus such as Times, British Broadcasting Corporation (BBC) and Cable News Network (CNN) with the accuracy rate being 38.28%. Due to the large number of informal texts in the real network, the above work does not perform well in the task of classifying irregular texts. In order to solve this problem, Neviarouskaya et al. first preprocessed the irregular texts such as informal abbreviations, emotional icons and grammatical errors in text, and then used semantic rules-based methods to divide all the emotions into nine categories which are joy, surprise, anger, sadness, fear, disgust, guilt, shame [61]. The accuracy of the emotional classification of this method can reach 72.6% in the datasets including diary blogs, fairy tales and news headlines. In summary, although the rule-based methods can obtain the classification result in a shorter time, and can add other rules to improve the accuracy of the emotion classification, when the amount of data is large, the maintenance of the rules is more complicated, and not easy to expand.

2.2.3 Machine Learning-based Methods

In the era of big data, capturing massive textual data from social media like Twitter and extracting valuable emotional information is the main task for machine learning-based methods.

Compared with long textual corpora such as fairy tales, textual data from social media is usually short, unstructured and full of emojis which are widely used to express different emotions. To solve this problem, Jonathon proposed an emotion classification method based on emoji which extracted a set of specific emotion symbols from textual corpus, and employed this sample set to train the classification model [62]. The classification results were tested separately on data from finance, Mergers and Acquisitions and the mix of two themes which proved that emoji features are independent in term of themes. Based on this theory, Ouyang et al. proposed a convolutional neural network-based emotion classification architecture which uses word2vec to extract word vectors from text and takes them as input to convolutional neural networks [63]. In the experiment of commentary corpus [1], the model performed well in emotion classification tasks, and the accuracy rate reached 45.4%.

In summary, the methods based on machine learning is superior to the dictionary-based and rule-based methods in accuracy, but the requirement of the quality of the sample data is strict, and it takes massive time and labor to mark the corpus.

2.2.4 Multi-label methods

Above emotion classification methods can only detect one emotion in a sentence. However, a sentence may express multiple emotions. In order to more accurately grasp the emotional information expressed in text, methods based on multi-label emotion classification were carried out.

Although the complexity and difficulty of classification methods based on multi-labels are relatively high, more and more attention is being paid on this area. Buitinck et al. proposed a multi-label emotion detection system for film reviews [64]. This method first marks a sentence as a subset of the preset emotional label set by word bag and chapter features, and then uses one-vs.-rest SVM and Random K-Labelsets (RAKEL) methods to classify the sentence separately. Multi-label classification is more complicated than the single-label classification where the evaluation criteria includes not only accuracy but also Hamming loss (HL), Average Precision (AP), and One Error (OE). RAKEL gets the better result which is 84.1% for average accuracy, 11.2% for HL, 89.8% for AP, and 24.6% for OE. In 2016, Zhou et al. proposed a

[1] https://www.rottentomatoes.com

method based on emotional distribution [65]. This method designed a mapping function from sentence to emotion distribution to describe multiple emotions and their respective intensities, and introduced Plutchik's "Wheel of Emotions" theory to improve the accuracy of emotion detection. The average accuracy of this method in the Chinese blog Ren-CECPs corpus is 66.54% with HL being 17.72%, 64.10% for AP, and 52.39% for OE.

2.2.5 Human Need Detection

Human need detection, which aims at analyzing human psychological needs, did not received much attention yet. Based on the theory that several types of human needs are universal and directly influence consumer purchase behavior, Yang et al. investigated customer behaviour and product categories to predict customers' needs [66]. However, the model they used is constructed only based on marketing practice and associated with products and services properties. In order to develop an annotation framework to analyze social media textual data for basic needs concepts, [22] proposed a multi-layered psychological-based reference model to assess citizen needs during any event at any time. In fact, the human needs detection model we employed in our platform is based on this work.

2.3 Human Affective State Analysis Platforms

Based on the above works, some social media analytics platforms have been constructed.'Sentiment' is a text analysis system developed by Corpora company [2] which can be used to judge the sentiment type towards a political party's policy or some product in texts. It assists governments and decision-making departments of some large companies to fully understand public's views on them. Fujitsu [3] developed a customer comment evaluation system. This system can extract comments of companies and their products from blogs and forums in Chinese, Japanese, and English, and make an overall evaluation based on the sentiment analysis of the comments. It can also present positive and negative information of the company's product at any specific time with dashboards. Liu et al. developed Opinion Observer which can analyze customer feedback and compare the qualities of various products and then display the results in a visual way [67]. Wilson et al. developed Opinion Finder which can automatically identify subjective sentences related to target products and extract the sentiment information in the sentences [68]. Goonie company developed a product evaluation system using their own goonie senti-

[2]https://www.corporafinance.com/

[3]https://www.fujitsu.com/

ment metrics core components and some Chinese text processing techniques [69]. The system crawls related text information on Weibo (the most popular microblogging site in China) and generates detailed public opinion analysis reports. Through these reports, users can obtain feedback from netizens and evaluate the company's products. Hao Wang et al. developed a system to predict the outcome of the US election in 2012 [70]. It analyzed the sentiment of a large amount of online commentary information related to each presidential candidate during the election. Li et al. proposed City Digital Pulse (CDP), a cloud-based heterogeneous data analysis platform [71]. The system collects geo-tagged data from Twitter and Instagram and conducts sentiment analysis on these data. Therefore, it allows users to investigate changes of human affective states in different cities. In 2019, Daniel et al. presented a multimodal affect and context sensing platform which is designed to enable easy prototyping of novel computer interfaces that sense, respond and adapt to human emotion [72]. To the best of our knowledge, there is no human psychological needs analysis platform based on social media. Therefore, we propose an interactive and scalable human needs analysis platform to solve this problem.

2.4 Affect-aware City

A smart city aims to improve urban functions and provided services, being often perceived as a living urban fabric, in which connected urban citizens, acting as active sensors, have the capacity to contribute even more efficiently to the spatial intelligence of cities. This "immaterial" dimension is related with the need that smart cities have to assess their citizen's feelings, perception and well-being, giving rise to an affect-aware city. Anton et al. discussed smart, playable, and affective cities from the viewpoints of how a city can be perceived and experienced by its citizens and how the city perceives its citizens with the assumption that sensors, actuators and computing systems are embedded in urban environments at the 2018 ArtsIT conference [73].

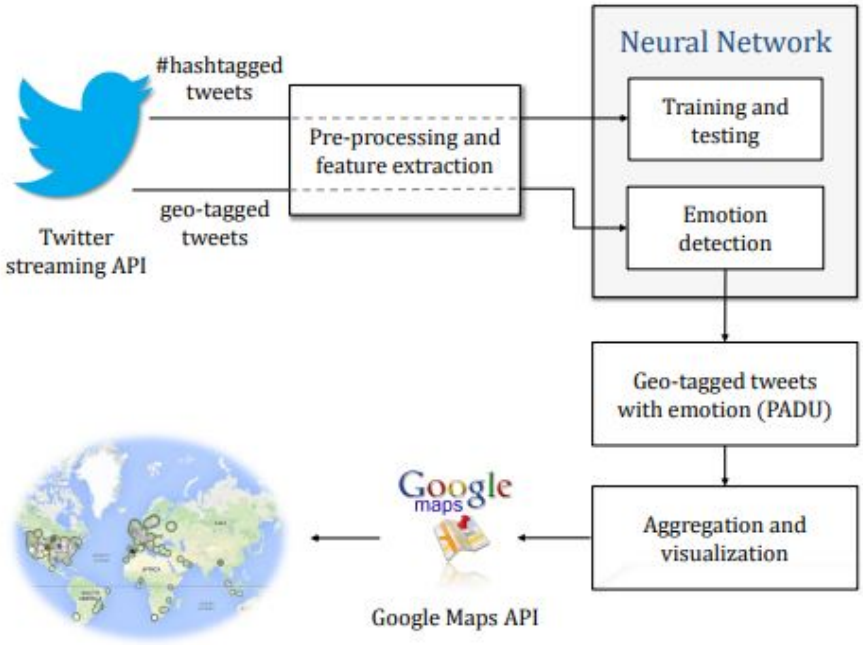


Figure 2.5: Architecture of the proposed system to detect and visualize emotions from geo-tagged Twitter messages [74].

In order to create affect-aware cities, a large number of affect-aware systems are needed. Landowska et al. proposed a framework that takes into consideration student emotional states and makes affective interventions [75]. The authors also provided definitions of 'affect-aware systems' and 'affective interventions' and describes the concept of the affect-awareness framework. Based on these conceptions, Thakur et al. proposed an algorithm that can effectively categorize the underlining emotion of an activity as positive or negative and then classify it amongst the six basic emotions - sadness, fear, happiness, anger, disgust and surprise [76]. This algorithm can be used to improve the user experience by optimizing the services of systems. The results presented show that the proposed model would help developers to understand user experiences in the form of user's emotions in any scenario. Oliveira et al. presented an Ambient Geographic Information (AGI) approach to assemble geo-tagged data from Twitter, Flickr, Instagram and Facebook related with people's perception and feelings regarding Lisbon (Portugal), and therefore characterize its emotional dimension, by comparing these subjective observations with objective measurements [77]. They employed socio-demographic statistics, questionnaires and data retrieved from biometric sensors. Results show that it is capable to in-

terpret and harnessing the emotional states of its citizens. However, this project needs a lot of manual work to gather and analyze information. To solve this problem, Guthier et al. proposed a system which could automatically detect public affections and visualizing them on a map using machine learning techniques [74]. The architecture of this system is shown in Figure 2.5. We can see that it has two phases: the offline training phase and the online phase. In the offline training phase, they collected raw tweets from Twitter using emotion hashtags like #happy, #sad and #fear. The tweets with the same hashtag can be used as a labelled data set for later supervised learning. Then they extracted features of the text and put them into a neural network to train and test the model. In the online phase, the geo-tagged tweets are feed to the trained model to detect emotion and the results are visualized on a map.

Chapter 3

System Design and Implementation

In this chapter, we first introduce the system architecture design. Then we illustrate how each module of the system is built in detail and how they interact with each other. The system implementation and security assurance system are described in the end.

3.1 Architecture Design

As shown in Figure 3.1, our proposed human needs analysis platform consists of four main parts: data collection module,data storage module, data analysis module and data visualization module. The four parts interact with each other and provide users with a thorough human needs analysis based on their queries. The whole process of how data flows in the system is illustrated as follows: Our crawlers in the data collection module gather raw data from a popular social network website Twitter. Then the data is fed into our human need detection model in the data analysis module before stored into the database. When a user enters a query through the user interface, they will get all the related items in the database by the index system of the data storage module and a comprehensive human needs analysis of these items is then presented and depicted in the data visualization module. The brief introduction of each module is described as follows:

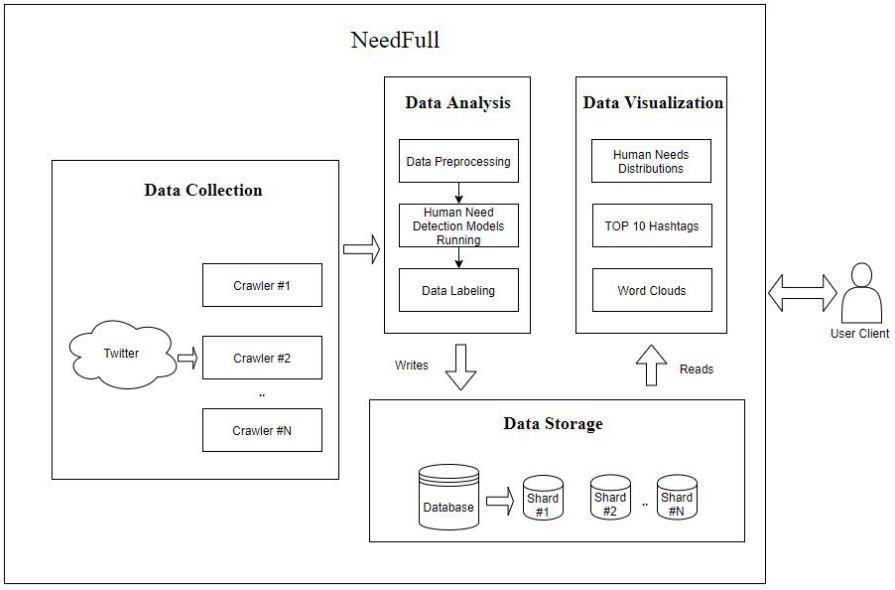


Figure 3.1: Architecture of NeedFull platform

- **Data collection module** is used to acquire online data. In NeedFull, we design and develop a two-layered web crawler to quickly crawl the original public tweets on Twitter. The first layer is built on the official Twitter API to obtain recent tweets and the second layer is a web crawler which can be used as a supplement in case historical data is required.

- **Data storage module** is designed to ensure the speed of data reading and writing. It also strengthens the security of the data and makes the system prepared for subsequent data analysis section and data visualization section.

- **Data analysis module** is employed to automatically detect human needs. When a tweet is fed into the data analysis module, it will go through some procedures including data preprocessing, feature extraction and human need models running.

- **Data visualization module** is designed mainly to provide a user-friendly human-computer interaction interface to display the human need detection results. In order to provide some insights about why people have those psychological needs, we list the top 10 most used hashtags and the most significant and frequently used words among the tweets.

3.2 Data Collection

There are there main ways to obtain the original tweets from Twitter:

- Official API interface
- Official or other Twitter data sets
- Web crawlers

The original tweets data is stored on the official server. We can obtain the required data by accessing the official data interface such as the Search API and the Streaming API provided by Twitter. Therefore, we do not need to spend time and effort to store data in our local data base if we only need recent tweets. Moreover, the tweets are stored in JSON format with clear structure, there is no need to clean up non-standard data. However, this type of interface is generally not open to the outside world or only a small number of interfaces can be used to obtain incomplete data access, and the data interface changes frequently. If the interface changes or upgrades, the source of the program needs to be modified. The access rights of the interface may also change at any time. Therefore, the interface data that can be accessed may not be obtained after a period of time which would cause great trouble to our system.

Official or other Twitter data sets store relevant data generated on the Twitter platform within a certain period of time. This type of data has been processed and packaged, so the data is clearly structured. Moreover, the data is stored locally and can be used at any time without worrying about data loss or inaccessibility. However, this type of data set only contains data for a certain period of time, when the data set is released and can be downloaded, the data has lost its timeliness which is fatal to public psychological needs analysis. Therefore, this type of data sets is not very useful.

The content of web pages on Twitter can be crawled by web crawlers. Because crawling is in real time, the crawled data is valuable and the information is complete. However, all data of the entire HTML page is crawled, it contains a lot of interference data. Moreover, most social networking platforms have anti-crawling measures which makes the crawling process extremely difficult.

The above three tweet data acquisition methods have their own advantages and disadvantages. Regarding the high demand for real-time and completed raw data of our system, we decided to use a two-layered collector for raw data acquisition. The first layer is built on the official Twitter API to obtain recent tweets and the second layer is a web crawler which can be used as a supplement in case historical data is required.

3.2.1 Recent tweets collector

Our recent tweets collector is built on the official Twitter Search API which allows developers to collect tweets published in the past 7 days. In order to obtain the access to use the API provided by Twitter, we need to obey OAuth which is an open protocol to allow secure API authorization in a simple and standard method from desktop and web applications [1]. How we obtained the developer access token is shown in Figure 3.2.

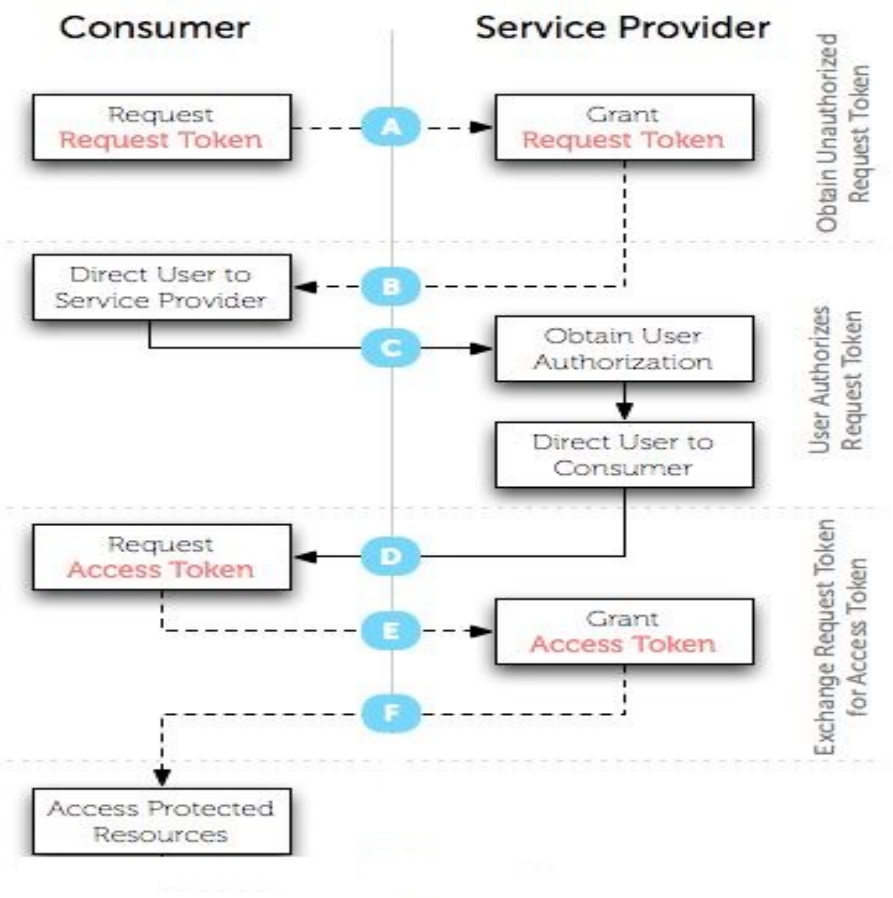


Figure 3.2: Applying for access to Twitter API

- A: A consumer requests a Request Token;
- B: The Service Provider grants a Request Token;
- C: A consumer directs users to the service provider;

[1] https://oauth.net/

- D: After obtaining the user authorization, the service provider directs users to a consumer;
- E: A consumer requests an Access Token;
- F: The Service Provider grants an Access Token;

After obtaining the access to the Search API, we can use it to collect any recent tweets published in the past 7 days. Since users of our system would be allowed to input queries of interest with keywords,dates and locations, we need to specify these three parameters in the Search API. We can see from Figure 3.3, Twitter's Search API continuously delivers new responses to REST API queries over a long-lived HTTP connection. Receiving updates on the latest Tweets matches a search query, stay in sync with user profile updates. The search process gets the input from Tweets and performs any parsing, filtering, and/or aggregation as needed before storing the results to a data store. The HTTP handling process queries the data store for results in response to user requests.

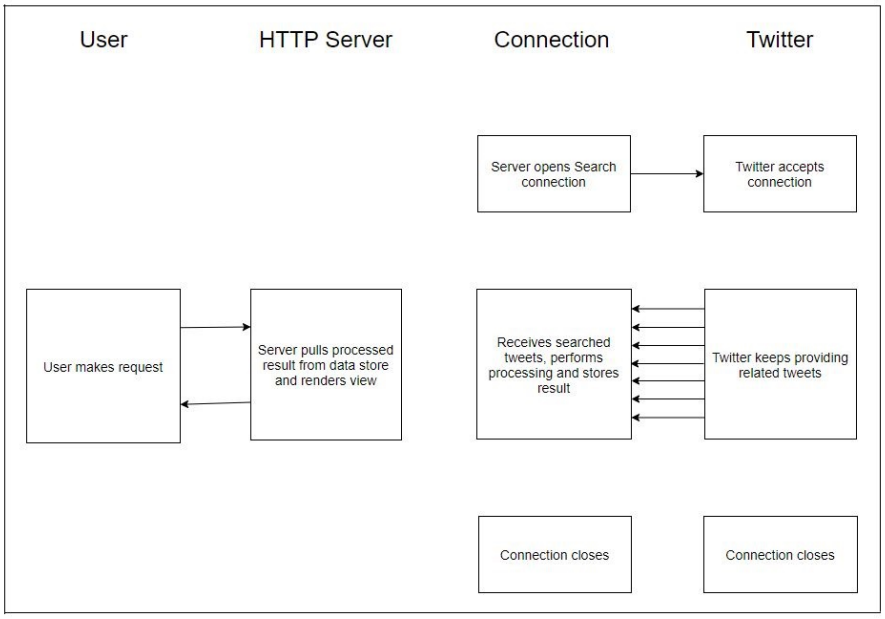


Figure 3.3: Twitter's Search API

3.2.2 Historical tweets collector

Our historical tweets crawler is built on Scrapy, an open-source web-crawling framework [2]. As shown in Figure 3.4, the crawler consists of five significant components: Engine, Scheduler, Downloader, Spider and Item Pipeline. Engine is responsible for controlling the data flow between all components of the system, and triggering events when certain actions occur. Scheduler receives requests from the engine and adds them to a queue for feeding them later when the engine requests them. Downloader is responsible for fetching web pages and feeding them to the engine which, in turn, feeds them to the spiders. Item Pipeline is responsible for processing the items once they have been extracted by the spiders. Here is the process of how it works:

1. The Engine gets the initial Requests to crawl from the Spider;
2. The Engine schedules the Requests in the Scheduler and asks for the next Requests to crawl;
3. The Scheduler returns the next Requests to the Engine;
4. The Engine sends the Requests to the Downloader;
5. Once the page finishes downloading the Downloader generates a Response (with that page) and sends it to the Engine;
6. The Engine receives the Response from the Downloader and sends it to the Spider for processing;
7. The Spider processes the Response and returns scraped items and new Requests (to follow) to the Engine,;
8. The Engine sends processed items to Item Pipelines, then send processed Requests to the Scheduler and asks for possible next Requests to crawl;
9. The process repeats (from step 1) until there are no more requests from the Scheduler;

[2] https://scrapy.org/

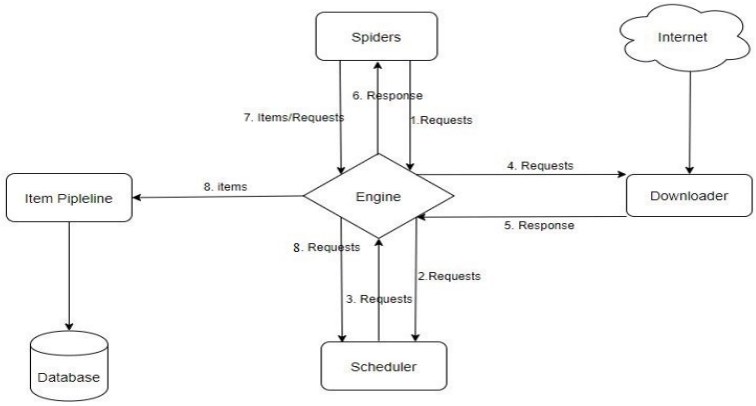


Figure 3.4: Architecture of NeedFull crawler for historical data

As the web page downloaded by the historical crawler is a HTML document which can be represented by the Document Object Model (DOM), in order to obtain the content we want, we need to parse the DOM structure of the web page first. Figure 3.5 shows a simple example of a DOM. If we can identify the location of the data that we want to collect, we can extract it easily.

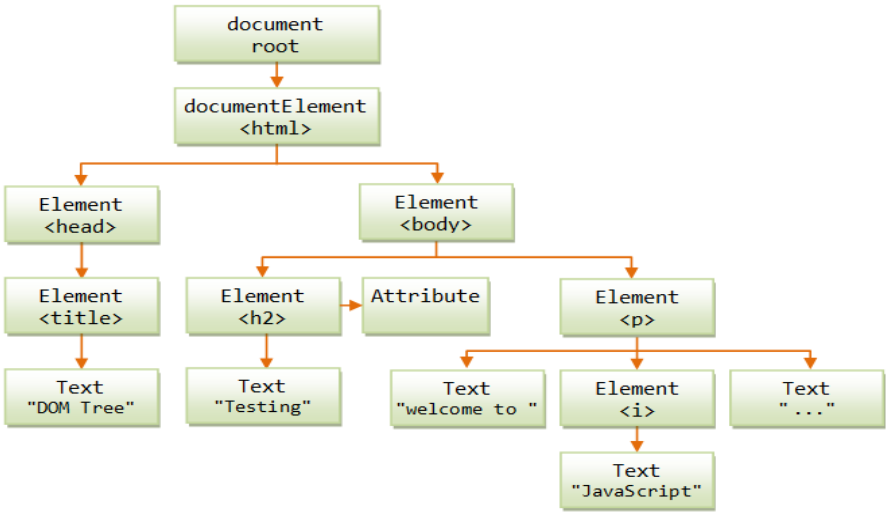


Figure 3.5: An example of DOM

3.3 Data Analysis

The framework we employed to automatically detect human needs in the system consists of two main phases: the offline phase for training and testing shown in Figure 3.6 and the online human need detection and analysis phase. In the offline phase, we employ machine learning methods on a psychological tweets data set labelled by three psychologists to develop the need models, namely: Need Content Recognition (NCR) model, Need Type Identification (NTI) model, Need Satisfaction Level Measurement (NSM) model, Social Context Evaluation (SCE) model, and Life Aspect Identification (LAI) model. The meaning of each model is explained as follows: Layer 1 (Need Content Recognition model) is designed to recognize need content in the post. If need content exists, the post will go to Layer 2 (Need Type Identification model), 3 (Need Satisfaction Level model) to identify the need type (relatedness, competence and autonomy) and its satisfaction level (satisfied, dissatisfied and neutral). Layer 4 (Social Context Evaluation model) is constructed to assess the quality of individual's social context (supportive, unsupportive and Not clear) and Layer 5 (Life Aspect Identification model) identifies the aspect of life that involves in the need experience. In the online phase, we collect data from Twitter and feed them into the trained model to obtain psychological results for each need model. Then the analysis of the results is presented. We introduce the methodology of both phases in this section.

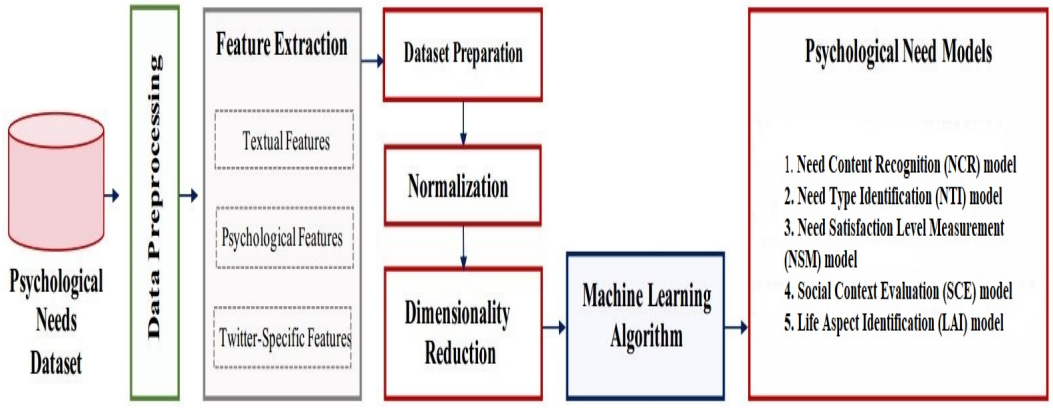


Figure 3.6: Offline training and testing phase of the human need detection framework[78]

3.3.1 Data Preprocessing

The tweets we download from Twitter are stored in JavaScript Object Notation (JSON) format where each tweet consists of dozens of key-value pairs. Since we only focus on analyzing the textual content, according to the JSON format, we obtain all the texts using the key 'text' which includes all the textual content posted by users. Before we start training these texts, we need to clean them up as there are a lot of useless information in the texts. Due to privacy issues, Twitter's usernames are removed and replaced by '@user'. This tag has no practical meaning for human need detection, so we get rid of all the '@user' in tweets by using a regular expression which defines a search pattern. For short words, like 'a', 'the', 'an', they almost exist in all the sentences and have no influence on the results. Therefore, we directly remove all these words by using 'stopwords', a NLP library Natural Language Toolkit (NLTK) which could automatically find all those meaningless words. And for words derived from the same word, like 'plays', 'playing', 'player', these words can be replaced by their original word 'play'. Therefore we employ 'stem', another library from NLTK to find these words and remove their suffixes. Because the tweets are full of emojis and emoticons which are widely used by people to express their feelings, we keep all the emojis and emoticons provided by Twitter and take them as a part of texts. Since online texts like tweets contain a large number of abbreviation especially for negation, like 'amn't', 'amnt', standing for 'am not'. We summarize all the common abbreviation expressions and correct them if they exist in the texts. One more thing we need to pay attention to is that people are used to using internet slang terms in tweets, so we sum up all the common slang terms, like 'LOL' or 'lol' standing for 'laugh out loud'. The last step for cleanup is to tokenize the remaining words in text which is a technique used to divide texts into words and make all the words serialized by assigning a number to each word.

3.3.2 Feature Extraction

Text feature extraction is the process of taking out a list of words from the text data and then transforming them into a feature set which is usable by a classifier [79]. In other words, a tweet needs to be converted to a vector which can represent the tweet. We believe that the human need detection result of a tweet is related to different features. Therefore, we investigate three categories of features: text-based features, psychological features, and Twitter-specific features.

For text-based features, we explore three widely used textual features extraction methods: Bag-of-words model (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and N-gram Language Model (LM).

BoW is a representation of text that describes the occurrence of words within a document. It is a simple and flexible way of extracting features from documents has been used with great success on prediction problems like language modeling and documentation classification. It involves two things: a vocabulary of known words and a measure of the presence of known words. BoW just keeps track of word counts and disregard the grammatical details and the word order. It is called a "bag" of words because any information about the order or structure of words in the document is discarded. However, it suffers from some shortcomings. The vocabulary requires careful design, most specifically in order to manage the size, which impacts the sparsity of the document representations. Discarding word order ignores the context, and in turn meaning of words in the document. Context and meaning can offer a lot to the model, that if modeled could tell the difference between the same words differently arranged ('this is interesting' vs 'is this interesting'), synonyms ('old bike' vs 'used bike'), and much more.

TF-IDF is a statistical measure that evaluates how relevant a word is to a document in a collection of documents. The tf-idf weight is composed by two terms: the first computes the normalized Term Frequency (TF), aka. the number of times a word appears in a document, divided by the total number of words in that document; the second term is the Inverse Document Frequency (IDF), computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears. TF-IDF was invented for document search and information retrieval. It works by increasing proportionally to the number of times a word appears in a document, but is offset by the number of documents that contain the word. Therefore, words that are common in every document, such as 'this', 'what', and 'if', rank low even though they may appear many times, since they do not mean much to that document in particular.

LM is a statistic model used for computing the probability distribution of a sequence of words, namely P(w_1, w_2, ..., w_n). where w_i is the ith word of a sequence with n words. According to the chain rule of probability:

$$P(w_1, w_2, ..., w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)...P(w_n|w_1, w_2, ..., w_{n-1})$$

In order to solve above equation, we introduce Markov Assumption assuming the appearing probability of the current word depends only on the previous n-1 words. In our case, we use bi-gram which assumes n as 2. Therefore, it can be written as:

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i|w_{i-1})$$

In order to make the conditional probability of the first word of a tweet meaningful,we add one indicator $< s >$ to the beginning of the original sequence and one indicator $< e >$ to the end of the sequence to make the bigram model able to model for sequences with any lengths. Therefore the original sequence turns into '$< s >, w_1, w_2, ..., w_n, < e >$'. We also introduce Laplace Smoothing to deal with out-of-vocabulary (OOV) words that they do not appeared in the training set but appear in the test set, resulting in zero probability calculated by the language model. In summary, n-gram takes the word order into consideration which is ignored by BoW and TF-IDF. Moreover, it employs maximum likelihood estimation, so the parameters are easy to obtain.

For psychological features, we explore two popular psychological related models: Linguistic Inquiry and Word Count (LIWC) and Linguistic Category Model (LCM). LIWC is designed to analyze the psychological meaning of words [80]. The LIWC dictionary contains 4500 words and word stems. Each is filed into one or more subdictionaries. Subdictionaries represent one of the 55 word categories through which LIWC compiles a text. For example, the word "cried" is part of "five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. Hence, if it is found in the target text, each of these five subdictionary scale scores will be incremented". When a matching word is extracted, the LIWC model calculates the percentage of total words that match each of the dictionary categories and forms the vector.

LCM is a linguistic classificatory approach that classifies verbs people use during any social events [81]. It consists of three linguistic categories: Descriptive Action Verbs (DAVs), Interpretative Action Verbs (IAV), and State Verbs (SV). Descriptive Action Verbs, are highly informative verbs that provide specific and concrete descriptions of actions during a short duration, such as hit, hold, jog. Interpretative Action Verbs, are verbs that describe enduring behaviors and events without describing the feature of the action, such as avoid, help, and attend. State Verbs are related to thoughts and affective states, such as think, understand, love, and hate. We count the frequency of all these there categories of verbs in a tweet and add it to the original vectors.

We believe that Twitter-specific features, such as widely used emojis and hashtags on Twitter, can make a big contribution to the prediction of human needs. We count the numbers of emojis and hashtags used in a tweet and explore the rule of the emoji's colors including: black, cream white, dark brown, moderate brown and pale. We also employ the sentiment of Emojis by employing the emoji sentiment lexicon created which consists of 751 emojis annotated manually by 83 human annotators as positive, negative and neutral [82].

3.3.3 Normalization

Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information. Because vectors used to represent tweets are composed of features from text-based, psychological and Twitter-specific three different aspects, the values in vectors have various ranges. However, features with bigger ranges may have bigger influence on the classification result. To reduce redundancy and dependency of data and change the values to a common scale, we employ normalization:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where x_{new} is the normalized value for x and it is in the range of [0, 1].

3.3.4 Dimensionality Reduction

Dimensionality reduction can be defined as the process of increasing the simplicity of a data set by reducing the number of random variables. Since our vectors contain a large number of various features, it may lead to Curse of Dimensionality where the accuracy of classifier goes lower as the number of features becomes larger. Therefore, we employ Principal Component Analysis (PCA) to perform feature selection. The main idea of PCA is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. The same is done by transforming the variables to a new set of variables, which are known as the principal components and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order. In this way, the first principal component retains maximum variation that is present in the original components. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal.

3.3.5 Evaluation of the Models

For the classification algorithm of each human need model, we employ a Support Vector Machine (SVM) which is a supervised machine learning model and offers very high accuracy compared to other classifiers such as logistic regression, and decision trees. The classification models' performance is evaluated using the metrics of Recall, Precision, F_{score} and Accuracy.

Psychological need recognition models	Accuracy %	R	P	F
Need Content Recognition (NCR) Model	79.13	0.791	0.82	0.77
Need Type Identification (NTI) Model	81.97	0.81	0.82	0.817
Need Satisfaction Level Measurement (NSM) Model	93.56	0.93	0.93	0.93
Social Context Evaluation (SCE) Model	75.70	0.75	0.76	0.75
Life Aspect Identification (LAI) Model	60.48	0.59	0.60	0.59

Table 3.1: Accuracy, Recall (R), Precision (P) and F_{score} for the human need models [83]

The results in Table 3.1 show that the need satisfaction level measurement model reaches the highest score with 93.56 % in accuracy followed by need type identification model with 81.97 % and need content recognition model with 79.13 %. For other metrics, the four models including need content recognition model, need type identification model, need satisfaction level measurement model, social content evaluation model reach around 80%. For life aspect identification model, all the results are not very good with around 60% in accuracy.

3.4 Data Storage

In order to satisfy the requirements of our system and save some storage space, we only extracted and stored important attributes such as id, text content, date, location and hashtags. Since these attributes are unstructured and the length may be indefinite. Therefore, if we use a traditional relational database for storage, we need to design multiple data tables and establish multiple relationship mappings. Meanwhile, considering that the stored attributes of tweets are not closely connected, and operations such as connection and aggregation will not be used, we decided to use not only Structured Query Language (NoSQL) database for storage. Unlike traditional relational databases that use the row and column structure as their basic unit, NoSQL databases use key / value pairs as their basic unit to store data. Each item of a NoSQL table can contain different data types, so there is no need to unify the data type when designing the data table, and we can flexibly store various types of data according to actual needs. Moreover, NoSQL databases remove relational operations such as connection and aggregation, and optimize the process of reading and writing data. As for social media analysis platforms such as NeedFull where large-scale data reading and writing operations are needed, the performance of NoSQL databases is better than the traditional relational databases.

In this thesis, we used MongoDB as our database. MongoDB is a cross-platform document-

oriented database program developed by MongoDB Inc [3] and some of its characteristics are shown as follows [84]:

- In MongoDB, you can search by field, range query and it also supports regular expression searches.
- It supports Master Slave replication.
- It has an automatic load balancing configuration because of data placed in shards.
- It supports JSON data model with dynamic schemas.
- Easy to administer in the case of failures
- You can index any field in a document.
- It can run over multiple servers. The data is duplicated to keep the system up and also keep its running condition in case of hardware failure.
- Stores files of any size easily without complicating your stack.

As our platform needs to support a small number of write operations but a large number of read operations, we built a replica set in Mongodb for the original database which has a primary node and multiple secondary nodes to scale out read operations. As Figure 3.7 shows, write operations such as inserting new data, updating or deleting previous data go to the primary database and read operations like querying go to the secondary databases. Moreover, a primary database could have multiple secondary databases and these databases can be synchronized by replicating the log in the primary database once there are new write operations in the primary database. We could add more secondary databases when the number of the data grows larger and the load balancer can easily find the relatively free one. This also makes the system easy to scale. Even when the primary database breaks down, we can still use these backup databases to keep the system running which significantly improves the security and robustness of our platform. Since the cost of performing a read operation in a single table with all the data is extremely high when the number of data grows extremely large, we partitioned the original table by building a table for each day of each city and created an index for every table. When a query with specified dates and cities enters our system, the load balancer can first allocate it to a relatively free secondary database node. Then the index will find the corresponding ids of items related to the query. Therefore, all the related items can be extracted fast.

[3] https://www.mongodb.com/

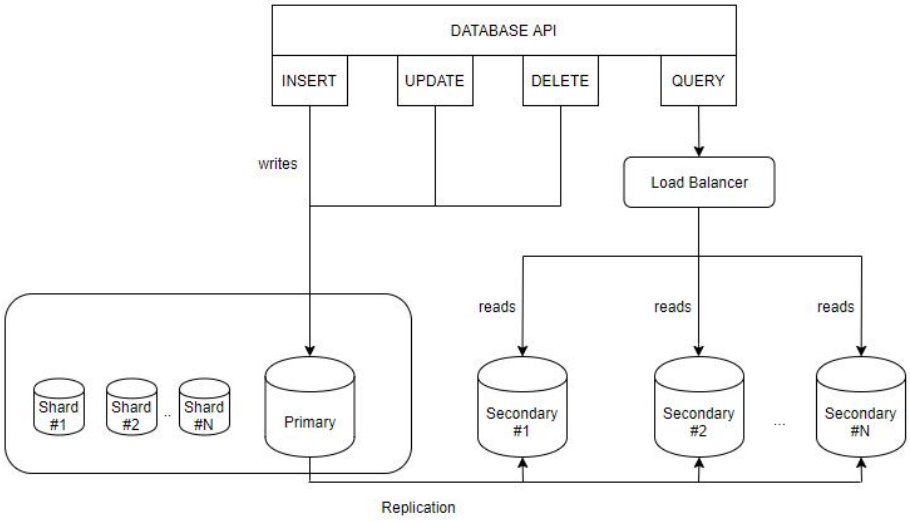


Figure 3.7: Design of database in NeedFull platform

3.5 Data Visualization

In order to provide users a thorough analysis of human affective states, the functions of our system includes: basic functions (registration, login, and editing personal info), display of human need distribution, visible analysis of human needs (top 10 hashtags and word clouds), and tweets examples. The use case diagram in the Unified Modeling Language (UML) is shown in Figure 3.8.

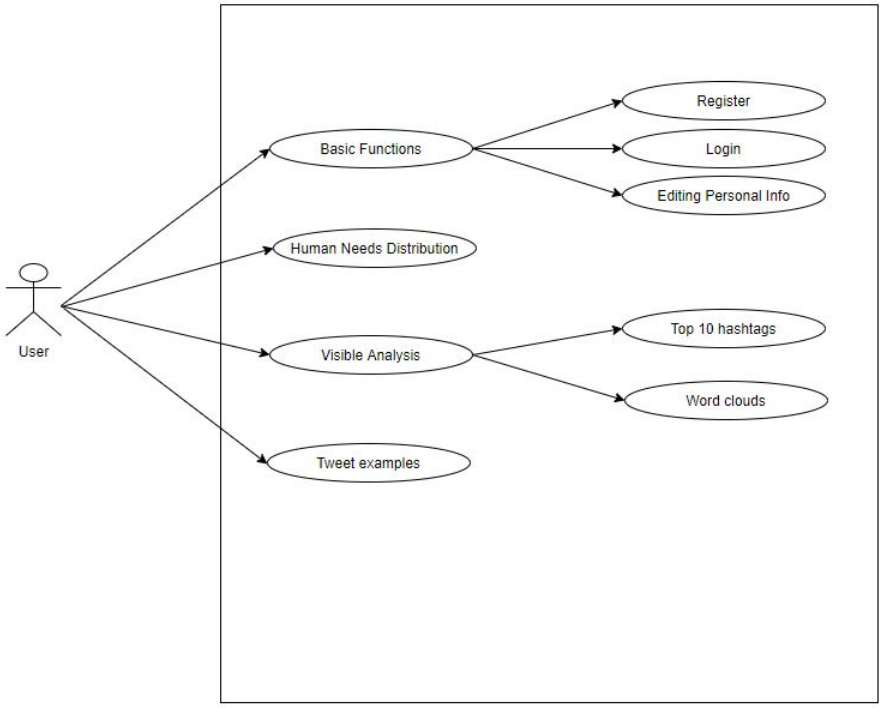


Figure 3.8: Use Case Diagram in UML

We designed a user-friendly interface to interact with users and visualize the analysis result. The interface is built on Chart.js which is a free open-source JavaScript library for data visualization. Users need to register for accounts in register page (Figure 3.9a) first and log in with their accounts in login page (Figure 3.9b). Once a query of interest with keywords, dates and locations is specified by users in search page (Figure 3.9c), they will enter the result page (Figure 3.9d) with a thorough human needs analysis. In the result page, it provides interactive dashboards of human needs distribution. Since hashtags are considered to be the user-driven method for categorizing tweets regarding specific topics, we believe these active topics related to the queries can explain why people have those psychological needs and why those needs are satisfied or not. Therefore. we select the top 10 most used hashtags in the related tweets and put them in the result page. Moreover, we explore the most significant and frequently used words among the tweets in order to determine the theme of the tweets and recognize the key aspects that were discussed. We plot word clouds generated by the significant words for each human need model. More details will be demonstrated in the next chapter with the analysis of public reaction to COVID-19 pandemic in New York State.

Figure 3.9: User interface of NeedFull: Register page (a), Login page (b), Search page (c), and Result page (d)

3.6 System Implementation

Our platform is developed on Django, a Python-based free and open-source web framework, which follows the model-view-controller (MVC) architectural pattern [4]. We can see from Figure 3.10, 'Model' handles the data representation and it serves as an interface to the data stored in the database itself, and also allows users to interact with the data without all the complexities of the underlying database. 'View' represents what you see on your browser. 'Controller' provides the logic to either handle presentation flow in the view or update the model's data. For example, it uses programmed logic to figure out what is pulled from the database through the model and passed to the view,also gets information from the user through the view and implements the given logic by either changing the view or updating the data via the model. Therefore, when users input a query through the search page in NeedFull, they will get the analysis results by a HTML web page.

[4]https://www.djangoproject.com/

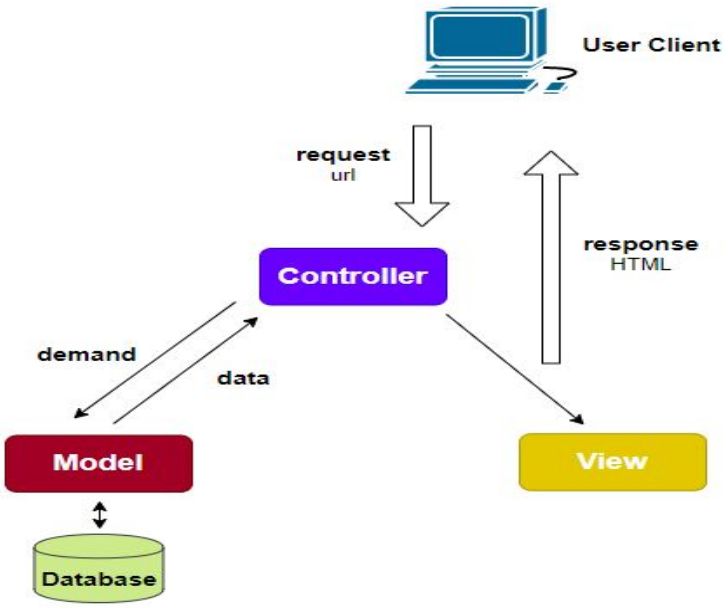


Figure 3.10: MVC Architectural Pattern of Django

The system development environment is divided into two parts: the hardware development environment and the software development environment. The entire system is constructed on a PC in Multimedia Communication Research (MCR) lab from University of Ottawa. The configurations of the machine are shown in Table 3.2:

The software development environment mainly includes tools used for front-end development, back-end development and data storage. Specific software and version information are shown in Table 3.3.

ITEM	VALUE
Operating System	Microsoft Windows 10 Enterprise x64
Processor	Intel Core i7-9700k CPU
Number of cores	8
Random Access Memory(RAM)	32.0 GB
Hard Disk	2 TB

Table 3.2: Hardware development environment

ITEM	Version
Django	Ver 2.2.5
Python	Ver 3.6.4
HTML	Ver 5
CSS	Ver 3
JavaScript	Ver 6
MongoDB	Ver 4.0.10

Table 3.3: Software development environment

3.7 Security Assurance System

Software security assurance is a process that helps design and implement software that protects the data and resources contained in and controlled by that software [85]. Software is itself a resource and thus must be afforded appropriate security . In our platform, the security assurance system mainly includes the following parts:

3.7.1 Operating system security

This system runs on the Windows operating system. We locked unnecessary system users and groups, disabled other irrelevant services to ensure the security of the operating system.

3.7.2 Data security

Data security refers to the process of protecting data from unauthorized access and data corruption throughout its lifecycle [86]. In order to protect the data of our platform, users with unauthorized access do not have access to the database. At the same time, all data on the database is backed up regularly.

3.7.3 Network security audit

We have recorded in detail the log information generated by our NeedFull platform on the windows system. When an abnormal situation occurs in the system, we can easily find the cause of the problem according to the log information.

Chapter 4

Human needs analysis of COVID-19 pandemic

4.1 Human needs analysis of COVID-19 pandemic in New York State

We inspected the reaction of people in New York State to the ongoing worldwide COVID-19 pandemic. The first case of COVID-19 in New York was confirmed on March 1. As of April 6, 2020, there have been over 130000 confirmed cases in the state, and of those 4159 people have died with the highest number of confirmed cases of any state in the United States.[1] To control the spread of the coronavirus, the state has ordered nonessential businesses to shut down, banned gatherings of any size and mandated that people stay six feet away from each other. People's daily life has been extremely influenced by the outbreak of the coronavirus which also leads to the tremendous amount of posts related to COVID-19 in social media platforms. To obtain as many related tweets as possible from Twitter platform, we used the hashtags such as #coronavirus, #COVID-19, #outbreak, #pandemic and #virus in our data collection module and ended up with a total of 154486 publicly accessible tweets from March 1 to April 5.

[1] https://www.cdc.gov/

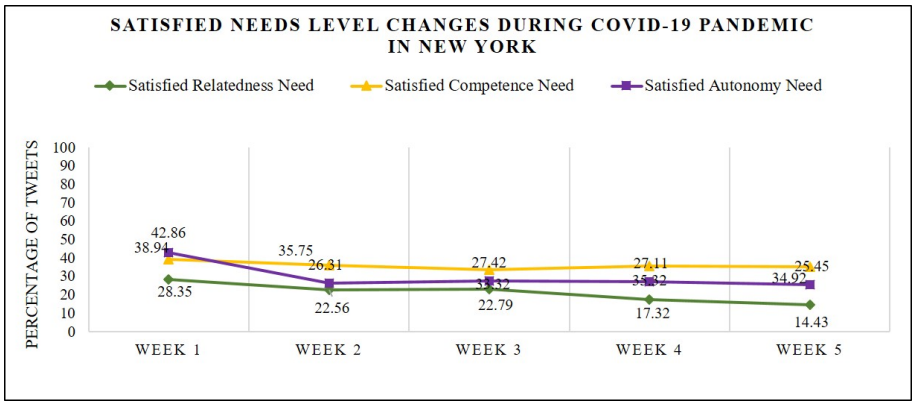


Figure 4.1: Measuring weekly change of satisfaction level during COVID-19 pandemic in New York State using NeedFull platform

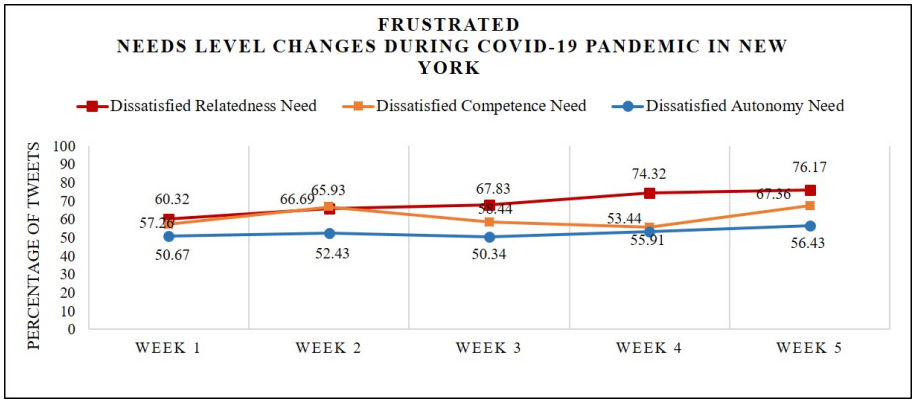


Figure 4.2: Measuring weekly change of frustration level during COVID-19 pandemic in New York State using NeedFull platform

4.1.1 Human needs analysis of COVID-19 pandemic in New York State in week 1

Figure 4.2 illustrates that in week 1, the first week when the first coronavirus case was confirmed, over half of the tweets expressed frustration for each psychological need: relatedness (60.32%), autonomy (50.67%), competence (57.26%). As shown in Table 4.1, people began to talk about it using the hashtags #CoronaOutbreak, #CoronavirusUSA, #CoronavirusChina (the first country where the corona virus outbreaks) and #CoronaVirusUpdate. Some measures proposed by the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) used to control the spread of the virus such as #WashYourHands and #PayAttention also caused a broad discussion. From the word cloud (a) in Figure 4.3, we can see

WEEK 1	WEEK 2	WEEK 3	WEEK 4	WEEK 5
#CoronaOutbreak	#CoronavirusPandemic	#pandemic	#pandemic	#CoronaVirusOutbreak
#CoronavirusUSA	#CoronaVirusUpdate	#nylockdown	#batflu	#pandemic
#CoronaVirusUpdate	#coronapocalypse	#CoronavirusPandemic	#Wuhan	#Trumpandemic
#WashYourHands	#outbreak	#socialdistancing	#CoronaVirusOutbreak	#Survive
#FakeNews	#CancelEverything	#ToiletPaperPanic	#CoronaVirusChallenge	#CoronaVirusChallenge
#CoronavirusChina	#NBAsuspended	#CoronavirusOutbreak	#China	#Quarantine
#Business	#SocialDistancing	#lockdown	#coronavirusupdates	#TrumpVirus
#CoronaVirusChallenge	#washyourhands	#CoronaCrisis	#LockdownNow	#stayhome
#PayAttention	#StockMarket	#ToiletPaperApocalypse	#Quarantine	#socialdistancing
#outbreak	#pandemic	#StayHome	#socialdistancing	#StayAtHomeSaveLives

Table 4.1: Top 10 most frequently used hashtags in tweets during COVID-19 pandemic in New York State using NeedFull platform

that words like ”CDC”, ”flu”, ”hand”, ”Trump”, ”work” and ”case” appear in a huge amount of tweets. Although there were not many confirmed cases by then, people started to worry about the outbreak in New York State since it has killed thousands of lives and extremely damaged the economy in China. The worries lowered the satisfaction level for each psychological need. Here are some examples collected from week 1:

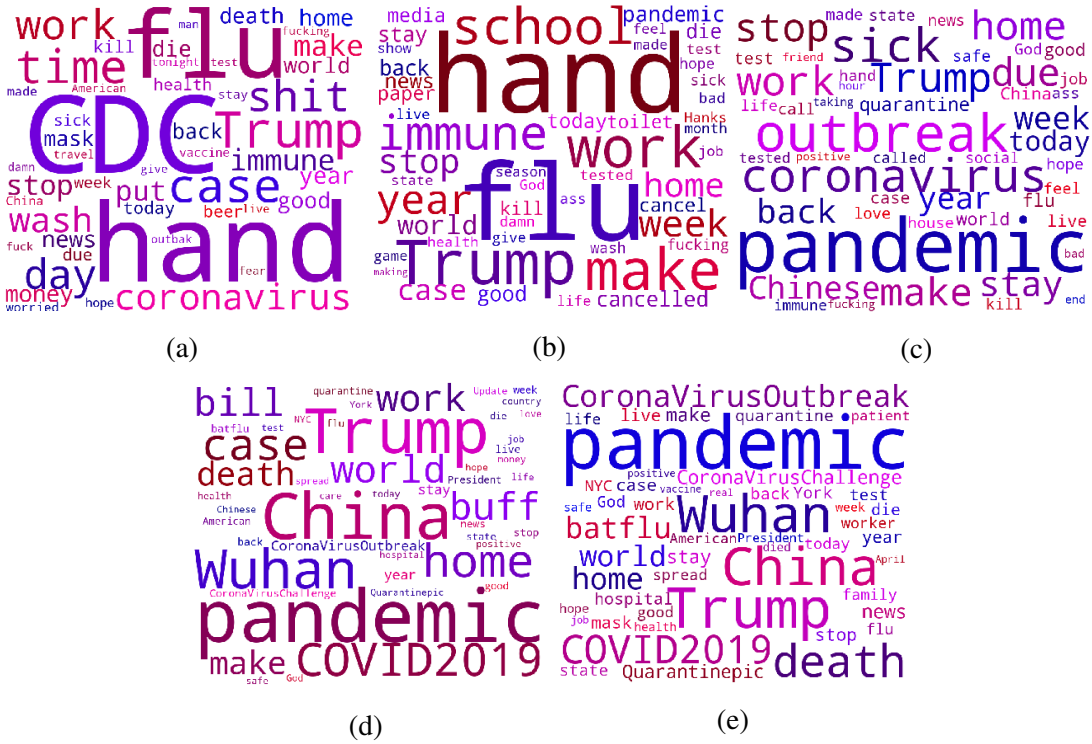
- I feel like this a great time to identify 1 or 2 people as your primary partners until this Corona Virus dies down. – Satisfied relatedness need
- I won’t be hanging my white friends till Corona virus over.-Dissatisfied relatedness need
- Yeah I get it, the flu has killed waaaay more people than the Corona Virus. But as a mother to young children I AM going to stress about the virus the same way I stress about the flu. If it has the possibility to KILL my children, I’m gonna worry. So shut the fuck up about it.-Satisfied autonomy need
- I think I got the corona virus and need to go home to isolate myself.-Dissatisfied autonomy need
- I have a better chance of listening to country on my own than getting the corona virus.-Satisfied competence need
- Why would we expect Trump and the GOP to do anything to protect our health from the corona virus or anything else when they have already depleted the EPA? They have rolled back 95 EPA protections endangering the environment and American lives.-Dissatisfied competence need

4.1.2 Human needs analysis of COVID-19 pandemic in New York State in week 2

In week 2, as can be seen from Figure 4.1, there is a noticeable reduction of satisfaction of residents' need. Compared to week 2, the satisfied relatedness need, autonomy need and competence need declined from 28.35% to 22.56%, 42.86% to 26.31%, and 38.94% to 26.31% respectively. Hashtags like #CoronavirusPandemic, #outbreak and #pandemic were widely used by residents in New York State shown in Table 4.1 as WHO has declared the coronavirus outbreak a pandemic on March 11. Protection measures such as #CancelEverything, #SocialDistancing and #washyourhands are still the center of topics since people realized the outbreak of coronavirus had begun in New York State. This can also been seen from the word cloud (b) in Figure 4.3 where. We can also find that some people pay much attention to the stock market and the National Basketball Association (NBA) using hashtags #StockMarket and #NBAsuspended. From online news, we know that, during this week, the coronavirus pandemic has had far-reaching consequences beyond the spread of the disease and efforts to quarantine it such as the crashed stock market and the suspended NBA season. We also noticed that the word "Hanks" showed up in the word cloud (b) in Figure 4.3. This is because the famous American actor Tom Hanks confirmed he and his wife had been diagnosed with coronavirus on March 11 which has aroused widespread concern. The impact of the coronavirus on people's lives was gradually increasing and people shared their feelings and worries about it. Some examples are listed as follows:

- BREAKING NEWS: Peter Griffin has tested positive for Corona Virus. Send prayers for him and his family. – Satisfied relatedness need
- I'm actually terrified for my kids to get this corona virus going around! Especially my baby because she's so young! -Dissatisfied relatedness need
- Can't wait to see what happens after the corona virus pandemic is over. -Satisfied autonomy need
- Looks like I may have lost my job thanks to the corona virus. -Dissatisfied autonomy need
- Can't believe stony brook is switching to all online classes in preparation for the corona virus that's so crazy but smart. -Satisfied competence need

- We do not have access to Corona tests, these numbers are not accurate. Not only is prevention and treatment the key to this PANDEMIC (worldwide) but knowledge is power. You are failing our children and our city. #besmart. -Dissatisfied competence need



(a) (b) (c)

(d) (e)

Figure 4.3: Word cloud generated from tweets during COVID-19 pandemic in New York State using NeedFull platform: WEEK 1 (a), WEEK 2 (b), WEEK 3 (c), WEEK 4 (d) and WEEK 5 (e)

4.1.3 Human needs analysis of COVID-19 pandemic in New York State in week 3

In week 3, there is no noticeable change of frustration and satisfaction level for each need with 67.83% of the relatedness, 58.44% of the autonomy and 50.34% of the competence being dissatisfied and 22.79% of the relatedness, 27.42% of the autonomy and 33.32% of the competence being satisfied shown in Figure 4.1 and Figure 4.2. It is easy to see that some new topics appear like #nylockdown and #lockdown from Table 4.1 as New York Governor Andrew Cuomo announced a strict lockdown including shutting down nonessential businesses

and ordering all 40 million state residents to stay at home on March 20. These measures to battle the spread of the COVID-19 have caused broad panic among residents and made them rush to stock up essential items such as toile paper, canned food, water bottles and pasta. This also explains the appearance of the hashtags #ToiletPaperPanic and #ToiletPaperApocalypse in Table 4.1. We expect these lockdown measures would bring a distinguishable increase in frustration level. However, deeper analysis shows frustration level did not change that much. It seems buying essential items and getting ready for staying at home for a long period eased people's anxiety in a way. Examples are posted in week 3:

- can't wait for my kid to come home one day and show me he's reading " I Survived: The Corona Pandemic of 2020". – Satisfied relatedness need
- My best friend can't even come to my "closest friends only" baby shower because of the stupid corona virus. -Dissatisfied relatedness need
- after the corona virus calms down, i will start living my best life with or without the attendance of others. -Satisfied autonomy need
- So if this corona virus thing keeps getting out of hand. I'll be quitting my job and I'm going quarantine myself at James Franco's house. -Dissatisfied autonomy need
- I told myself I wasn't going to go out this weekend. Thanks corona virus for helping me accomplish my goals. -Satisfied competence need
- i'm so over this corona virus. i just want to participate in my extracurricular and then graduate. every other high schooler except c/o 2020 had that chance. what did we do:(-Dissatisfied competence need

4.1.4 Human needs analysis of COVID-19 pandemic in New York State in week 4

From Figure 4.2, we can see that, in week 4, there is a slight increase in frustration level for each need compared to week 3: relatedness (74.32%), autonomy (53.44%) and competence (55.91%). As Table 4.1 shows, people began to use some new hashtags such as #batflu which is considered by some people as how the first case in China got infected, #Wuhan (the first city where the outbreak of coronavirus happened) and #China. The word cloud (d) in Figure 4.3 also demonstrates that the most frequently used words are "China", "Wuhan", "Chinese" and "pandemic" as president of the United State Donald Trump kept using "Chinese Virus"

instead of coronavirus or COVID-19 which connected the virus with China. However, the sustainable growth of frustration level denotes President Trump's comments did not satisfy people's psychological need though people talked more about China. Here are some tweet examples:

- Wonderful gesture James! Hollywood are you listening? Now would be a great time to show your appreciation for your success by publicly donating to a cause to defeat Corona Virus! I might even go to one of your movies. – Satisfied relatedness need
- Phew! even Chinese are not shopping here in New York their own people why are people are so judgmental and look at every Chinese as if they have corona virus which is not true! I have a lot of Chinese friends and so what? – Dissatisfied relatedness need
- Putting a little more time into cooking now that I am home more often during the Corona Virus Pandemic. These are photographs of me and my healthy lunch today. Remember good nutrition in the mist of staying home (avoid unhealthy snacks)! -Satisfied autonomy need
- Corona virus is so trash.... I was supposed to start swimming classes this weekend. -Dissatisfied autonomy need
- My move in date for my first apartment keeps getting pushed back...I'm about to find a cure for this corona virus myself. -Satisfied competence need
- I have lost my job a week before the corona virus shut down and applied for unemployment and was denied immediately. I then found a new job being an assistant teacher at a daycare but once the mayor ordered for schools to close then that meant my job closed so I couldn't start. -Dissatisfied competence need

4.1.5 Human needs analysis of COVID-19 pandemic in New York State in week 5

As the number of confirmed cases of coronavirus in New York State had gotten to over 130000 with 4159 people died in week 5, satisfaction levels, shown in Figure 4.1, reached the lowest so far where the percentages of satisfied relatedness, autonomy and competence are 14.43%, 25.45% and 34.92% respectively. From the topics people were talking about shown in Table 4.1, we can see that more and more residents started to blame Trump for this severe situation using hashtags #TrumpVirus and #Trumpandemic. From the online news, we can know that

Trump did not take coronavirus seriously when the outbreak of coronavirus just happened which makes people feel they are paying for his arrogance. In the meantime, the appearance of hashtags such as #Survive and #StayAtHomeSaveLives denotes that people worried their own safety more than ever. As the word cloud (e) in Figure 4.3 shows, some new words such as ”god”, ”death”, ”died”, ”week”, ”health” and ”hospital” were widely used by residents. Examples are shown as follows:

- With everything that is going on with the Corona virus. I was glad that my friend decided to check up on me and my family last night. We need all of the love and support at this moment. – Satisfied relatedness need
- Trump downplayed the corona virus for months. Called it the latest democrat hoax, said we have 15 cases that will soon be 0, how can you possibly defend this obvious mismanagement? -Dissatisfied relatedness need
- Good Morning! Its Monday, Rise and Grind! Don’t let that bitch Corona stop you from accomplishing something today. Big or small goal, today YOU got it! (Cute kitten to make monday better) #MotivationalQuotes #Covid-19. -Satisfied autonomy need
- I just spent a literal hour, calling different agencies. Local and state, trying to find a place where I can report a potential corona virus violation of the Governors order to do my part as a civilian and have it looked into and got nowhere. -Dissatisfied autonomy need
- This Corona Virus is just a setback I’m still striving to achieve my goals for this year. – Satisfied competence need
- The CDC has been warning us about a future pandemic we were unprepared for. Trump cut funding to CDC and scientific research into diseases in the past and while Corona Virus is not Trumps doing, his policies made it harder for us now. -Dissatisfied competence need

4.1.6 Overall human needs analysis of COVID-19 pandemic in New York State

The analysis result shows that, throughout this period, the most pronounced human need in these tweets is relatedness with 47.42%, followed by autonomy with 16.63% and competence

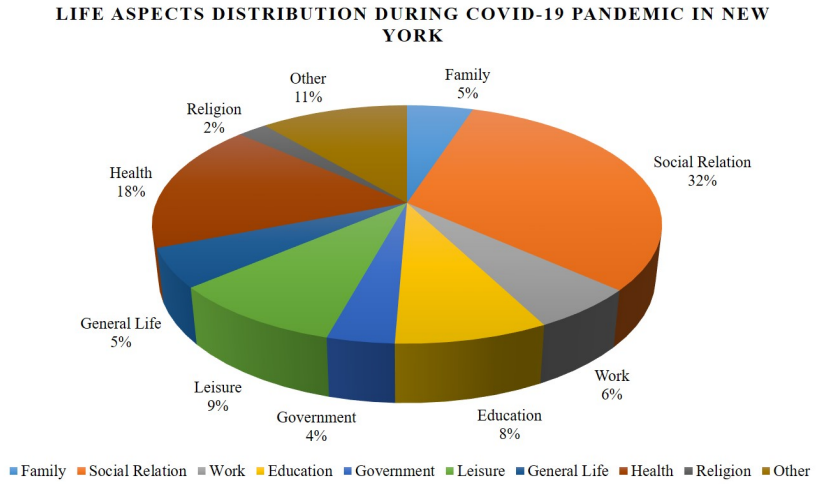


Figure 4.4: Recognizing life aspects distribution during COVID-19 pandemic in New York State using NeedFull platform

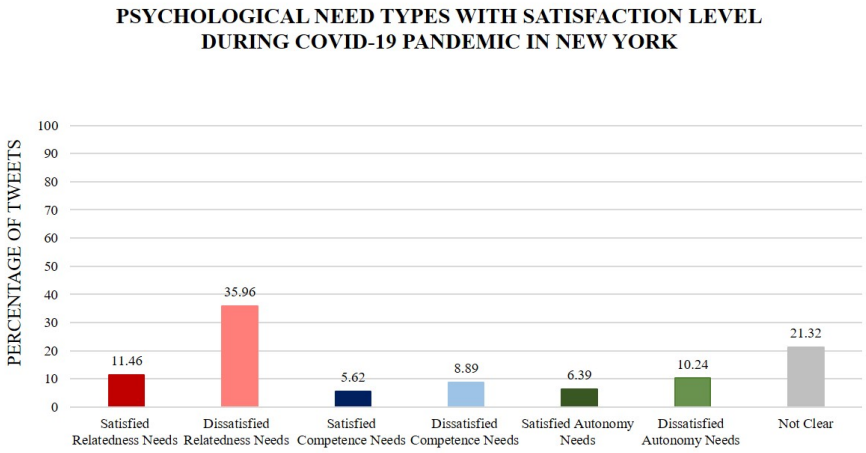


Figure 4.5: Recognizing human needs distribution during COVID-19 pandemic in New York State using NeedFull platform

with 14.51%. As can be seen from Figure 4.5, in general, the percentages of tweets expressing frustration are larger than those of tweets expressing satisfaction for each psychological need. For example, there are 35.96% of tweets revealing dissatisfied relatedness need while the percentage of satisfied relatedness is 11.46%. Figure 4.1 and Figure 4.2 demonstrate how the need satisfaction level changed weekly. We believe ongoing social topics play an important role in the changes of people's psychological needs. Therefore, we used the term frequency

(TF) measurement to acquire top 10 frequent hashtags for every week as hashtags are mainly used to denote specific topics of conversation in Twitter. Moreover, we studied the most frequently used words regarding to coronavirus and present them in word clouds where the size of a word shows how important it is in the discussion. Figure 4.4 shows the percentages of various aspects of an individual's life that people talked about concerning COVID-19. It is easy to see that the most pronounced life aspect is social relation with 32%, followed by health (18%), leisure (9%) and education (8%). It denotes people were eager to connect with their family and friends and they were worried about their health and education in the meanwhile. Examples for each life aspect are shown as follows:

- My mom is now scared of the Corona virus in NY so no vacation for me cause I'm not about to go alone! – Family
- We have many difference of opinions, even with our own friends, but when it comes to a common enemy, we have to put our difference of opinions aside and join our hands to fight together to extinct deadliest Virus than #Corona & save future. #Maridhas #Rajinikanth. – Social Relation
- I might go to work even if I get the corona virus. – Work
- My school just cancelled classes for next week due to the Corona Virus. #CoronaOutbreak. – Education
- Where can I get a free corona test Mr. President because your director of the CDC says I can get one but I've made several phone calls and I can't get one. – Government
- i don't know how i feel about planning a trip to vegas in may during this whole corona virus thing. – Leisure
- Can't wait till this corona virus is over so I can stop washing my hands every time I go to the bathroom. – Health
- All my friends. My birthday was Tuesday March the 10th, but I'm extending because of the Corona virus, so I'm accepting presents over the next two weeks until the 25th of March, just wanted to let everybody know. – General Life
- Everyone please pray for my family. My mom got the corona virus. – Religion

For social context analysis, Figure 4.6 illustrates how people in New York State evaluate their surrounding environment regarding the satisfaction level of their psychological need during this period. In general, the percentages of unsupportive social context are larger than those of supportive social context for every week and the difference between supportive social context (12.82%) and unsupportive social context (26.65%) reached the largest number in week 5. This confirms that people's general need were not satisfied over this period and the frustration level reached the highest in the last week shown in Figure 4.1 and Figure 4.2. As Figure 4.6 shows, there was a slight recovery in week 3 with 16.72% of the social context being supportive. This affirms our statement that buying essential items and getting ready for staying at home for a long period eased people's anxiety and made people have a better evaluation of social context. Here are some tweet examples:

- My group chat is having a very detailed conversation about the corona virus, I'm glad I'm friends with smart people.-Supportive social context
- FOR CORONA VIRUS stay home make yourself happy and do something I'm cooking every day good and healthy food, and I share to help you .The best time to learn anything for your self is now FOR CORONA VIRUS.-Supportive social context
- If my last semester at college is me taking online classes because of corona virus I will go into a depression.-Unsupportive social context
- My job really sent me home because they think I have corona virus I'm dead.-Unsupportive social context

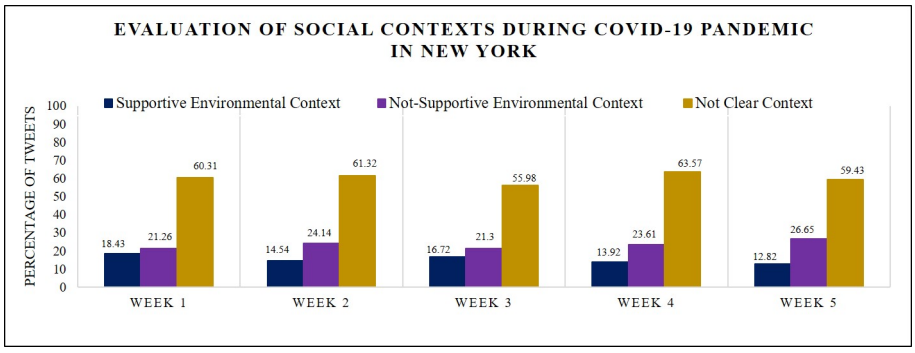


Figure 4.6: Identifying social context types during COVID-19 pandemic in New York State using NeedFull platform

4.2 Human needs analysis of COVID-19 pandemic in Ottawa, Toronto and Montreal

We also inspected the human needs of residents in Ottawa, Toronto and Montreal towards the ongoing worldwide COVID-19 pandemic. We focused on the period from March 1 to April 5 and got 98735, 11682 and 86921 tweets for Ottawa, Toronto and Montreal respectively. The human needs distribution for each city is shown as follows:

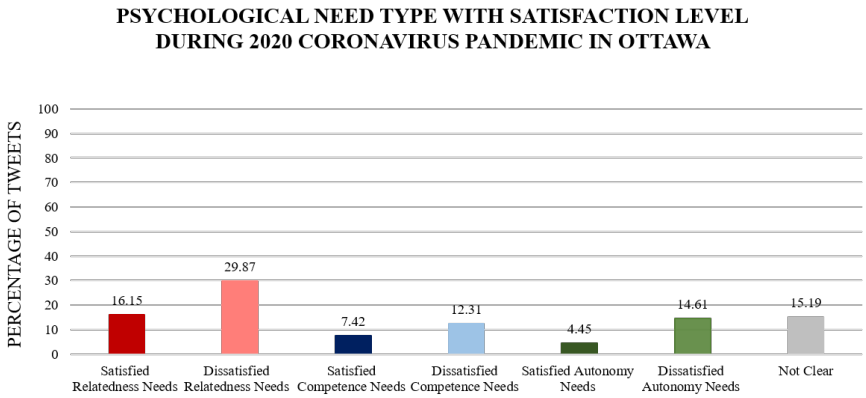


Figure 4.7: Recognizing human needs distribution during COVID-19 pandemic in Ottawa using NeedFull platform

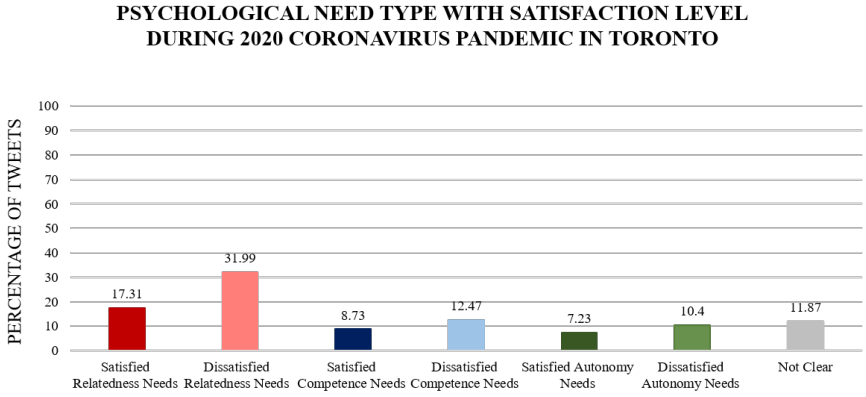


Figure 4.8: Recognizing human needs distribution during COVID-19 pandemic in Toronto using NeedFull platform

Figure 4.7, Figure 4.8 and Figure 4.9 show that the most pronounced human need during this period is relatedness (45.63%) followed by competence (21.37%) and autonomy (18.45%)

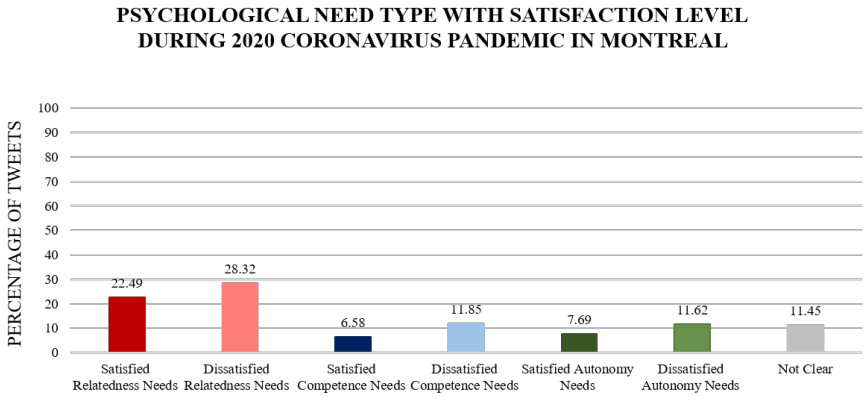


Figure 4.9: Recognizing human needs distribution during COVID-19 pandemic in Montreal using NeedFull platform

for these three cities. Moreover, in general, the percentages of tweets expressing frustration are larger than those of tweets expressing satisfaction for each psychological need for each city.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this work, we propose NeedFull, a tweet analysis platform, to help governments and municipalities to understand residents' reaction during critical times and dangerous situations. Different from the traditional social data analysis platforms, we focus on analyzing human psychological needs rather than sentiments by employing a multi-layered human need detection framework. The human need detection framework includes five need models, namely: Need Content Recognition (NCR) model, Need Type Identification (NTI) model, Need Satisfaction Level Measurement (NSM) model, Social Context Evaluation (SCE) model, and Life Aspect Identification (LAI) model. We explore three different categories of features: text-based features, psychological features, and Twitter-specific features when extracting features from the tweets and use multiple machine learning methods, such as Support Vector Machine, Principal Component Analysis to train these models. Results show that these models reached high accuracy: NCR (79.13%), NTI (81.97%), NSM (93.56%), SCE (75.70%), and LAI (60.48%).

NeedFull mainly consists of four parts: data collection module, data storage module, data analysis module and data visualization module. The four parts interact with each other to provide users with a thorough human needs analysis based on their queries. The whole process of how data flows in the system is illustrated as follows: Our crawlers in the data collection module gather raw data from a popular social network website Twitter. Then the data is fed into our human need detection model in the data analysis module before stored into the database. When a user enters a query through the user interface, they will get all the related items in the database by the index system of the data storage module and a comprehensive human needs analysis of these items is then presented and depicted in the data visualization module. In

data collection module, we design and develop a two-layered web crawler to quickly crawl the original public data on Twitter. The first layer is built on the official Twitter API to obtain recent tweets and the second layer is a web crawler which can be used as a supplement in case historical data is required. As our platform needs to support a small number of write operations but a large number of read operations, we built a replica set in data storage module for the original database which has a primary node and multiple secondary nodes to scale out read operations. Since a primary database could have multiple secondary databases, we could add more secondary databases when the number of the data grows larger and the load balancer can easily find the relatively free one. This also makes the system easy to scale. Moreover, these databases can be synchronized by replicating the log in the primary database once there are new write operations in the primary database. We employed the proposed platform to investigate the reaction of people in New York State and three big cities in Canada including Ottawa, Toronto and Montreal to the ongoing worldwide COVID-19 pandemic. The results show that their psychological needs are generally not satisfied due to the pandemic and the situation is getting worse as more people get infected with COVID-19.

5.2 Future Work

Human affective states analysis, especially human need analysis, is still in its infancy since it did not get enough attention from researchers and companies. As a tweet analysis platform to study human needs during critical times and dangerous situations, the research has achieved initial results, but it still has a long way to go. Future research in this thesis includes the following aspects:

- There are a lot of social media platforms which people are using to express their opinions, feelings and psychological needs. However, our data is all downloaded from Twitter. In order to enrich the database, we need to collect data from multiple social media platforms.

- Nowadays, more and more people would like to post images and videos on social media platforms. Therefore, focusing only on texts to analyze human needs is not enough. We are planning to explore the human needs analysis based on images and videos first and then add images and videos into our database for further study.

- NeedFull is a human need analysis platform. In order to provide more meaningful human affective states analysis, we need to include sentiment analysis and emotion analysis into

the platform.

Appendix A

Glossary of Terms

ABSA **A**spect-**B**ased **S**entiment **A**nalysis
AP **A**verage **P**recision
BBC **B**ritish **B**roadcasting **C**orporation
CNN **C**able **N**ews **N**etwork
CDC **C**enters for **D**isease **C**ontrol
CHI **CHI**-square
CDP **C**ity **D**igital **P**ulse
CBOW **C**ontinuous **B**ag-**O**f-**W**ords
DOM **D**ocument **O**bject **M**odel
DF **D**ocument **F**requency
EaLDA **E**motion-aware **L**atent **D**irichlet **A**llocation
GRU **G**ated **R**ecurrent **U**nit
HL **H**amming **L**oss
HNT **H**uman **N**eeds **T**heories
HTML **H**ypertext **M**arkup **L**anguage
IG **I**nformation **G**ain
LSA **L**atent **S**emantic **A**nalysis
LSTM **L**ong **S**hort-**T**erm **M**emory
ME **M**aximum **E**ntropy
MTV **M**odel-**T**emplate-**V**iew
MVC **M**odel-**V**iew-**C**ontroller
MAU **M**onthly **A**ctive **U**sers
MCR **M**ultimedia **C**ommunication **R**esearch

MI **M**utual **I**nformation
NB **N**aive **B**ayes
NER **N**ame **E**ntity **R**ecognition
NBA **N**ational **B**asketball **A**ssociation
NMT **N**eural **M**achine **T**ranslation
NoSQL **N**ot only **S**tructured **Q**uery **L**anguage
OE **O**ne **E**rror
PoI **P**oint of **I**nterest
RAM **R**andom **A**ccess **M**emory
RF **R**andom **F**orest
RAKEL **RA**ndom **K**-lab**EL**sets
RNN **R**ecurrent **N**eural **N**etwork
SO **S**emantic **O**rientation
SO-CAL **S**emantic **O**rientation **CAL**culator
SVM **S**upport **V**ector **M**achine
TF **T**erm **F**requency
TF-IDF **T**erm **F**requency weighted by **I**nverse **D**ocument **F**requency
UML **U**nified **M**odeling **L**anguage
WHO **W**orld **H**ealth **O**rganization

References

[1] J. E. Carlisle and R. C. Patton, "Is social media changing how we understand political engagement? an analysis of facebook and the 2008 presidential election," *Political research quarterly*, vol. 66, no. 4, pp. 883–895, 2013.

[2] A. H. Ali, "The power of social media in developing nations: New tools for closing the global digital divide and beyond," *Harv. Hum. Rts. J.*, vol. 24, p. 185, 2011.

[3] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *European conference on information retrieval*, Springer, 2011, pp. 338–349.

[4] A. Rapp, L. S. Beitelspacher, D. Grewal, and D. E. Hughes, "Understanding social media effects across seller, retailer, and consumer interactions," *Journal of the Academy of Marketing Science*, vol. 41, no. 5, pp. 547–566, 2013.

[5] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Information Fusion*, vol. 28, pp. 45–59, 2016.

[6] S. Tuarob and C. S. Tucker, "Quantifying product favorability and extracting notable product features using large scale social media data," *Journal of Computing and Information Science in Engineering*, vol. 15, no. 3, 2015.

[7] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith, "Part-of-speech tagging for twitter: Annotation, features, and experiments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, Association for Computational Linguistics, 2011, pp. 42–47.

[8] P. A. Grabowicz, J. J. Ramasco, E. Moro, J. M. Pujol, and V. M. Eguiluz, "Social features of online networks: The strength of intermediary ties in online social media," *PloS one*, vol. 7, no. 1, 2012.

[9] S. Tuarob and C. S. Tucker, "Automated discovery of lead users and latent product features by mining large scale social media networks," *Journal of Mechanical Design*, vol. 137, no. 7, 2015.

[10] A. Ceron and F. Negri, "The "social side" of public policy: Monitoring online public opinion and its mobilization during the policy cycle," *Policy & Internet*, vol. 8, no. 2, pp. 131–147, 2016.

[11] A. D. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014.

[12] M. Laroche, M. R. Habibi, M.-O. Richard, and R. Sankaranarayanan, "The effects of social media based brand communities on brand community markers, value creation practices, brand trust and brand loyalty," *Computers in Human Behavior*, vol. 28, no. 5, pp. 1755–1767, 2012.

[13] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[14] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Fourth international AAAI conference on weblogs and social media*, 2010.

[15] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining.," in *LREc*, vol. 10, 2010, pp. 1320–1326.

[16] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in *International conference on discovery science*, Springer, 2010, pp. 1–15.

[17] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceedings of the 23rd international conference on computational linguistics: posters*, Association for Computational Linguistics, 2010, pp. 241–249.

[18] A. Celikyilmaz, D. Hakkani-Tür, and J. Feng, "Probabilistic model-based sentiment analysis of twitter messages," in *2010 IEEE Spoken Language Technology Workshop*, IEEE, 2010, pp. 79–84.

[19] P. Chakriswaran, D. R. Vincent, K. Srinivasan, V. Sharma, C.-Y. Chang, and D. G. Reina, "Emotion ai-driven sentiment analysis: A survey, future research directions, and open issues," *Applied Sciences*, vol. 9, no. 24, p. 5462, 2019.

[20] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[21] M. Milyavskaya and R. Koestner, "Psychological needs, motivation, and well-being: A test of self-determination theory across multiple domains," *Personality and individual differences*, vol. 50, no. 3, pp. 387–391, 2011.

[22] R. Alharthi, B. Guthier, C. Guertin, and A. El Saddik, "A dataset for psychological human needs detection from social networks," *IEEE Access*, vol. 5, pp. 9109–9117, 2017.

[23] B. Guthier, R. Abaalkhail, R. Alharthi, and A. El Saddik, "The affect-aware city," in *2015 International Conference on Computing, Networking and Communications (ICNC)*, IEEE, 2015, pp. 630–636.

[24] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams engineering journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[25] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level sentiment classification: An empirical comparison between svm and ann," *Expert Systems with Applications*, vol. 40, no. 2, pp. 621–633, 2013.

[26] Y. Zhao, S. Dong, and L. Li, "Sentiment analysis on news comments based on supervised learning method," 2014.

[27] L. L. Dhande and G. K. Patnaik, "Analyzing sentiment of movie review data using naive bayes neural classifier," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. 3, no. 4, pp. 313–320, 2014.

[28] D. M. Mathews and S. Abraham, "Lexicon based document level sentiment analysis on the multilingual dataset," *Available at SSRN 3349025*, 2019.

[29] F. Liu, L. Zheng, and J. Zheng, "Hienn-dwe: A hierarchical neural network with dynamic word embeddings for document level sentiment classification," *Neurocomputing*, 2020.

[30] P. D. Turney, "Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2002, pp. 417–424.

[31] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[32] F. Li, M. Huang, and X. Zhu, "Sentiment analysis with global topics and local dependency," in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[33] Z. Li, X. Ma, and H. Xin, "Feature engineering of machine-learning chemisorption models for catalyst design," *Catalysis today*, vol. 280, pp. 232–238, 2017.

[34] X. Chen, Z. Liu, and M. Sun, "A unified model for word sense representation and disambiguation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1025–1035.

[35] J. Turian, L. Ratinov, Y. Bengio, and D. Roth, "A preliminary evaluation of word representations for named-entity recognition," in *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*, 2009, pp. 1–8.

[36] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[37] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[38] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext. zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[39] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[40] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in neural information processing systems*, 2014, pp. 2177–2185.

[41] Y. Goldberg and O. Levy, "Word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[42] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI Global, 2010, pp. 242–264.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[44] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," *arXiv preprint arXiv:1404.2188*, 2014.

[45] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[46] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[47] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[48] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[49] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[50] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[51] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," *arXiv preprint arXiv:1709.00893*, 2017.

[52] F. L. Cruz, J. A. Troyano, F. Enríquez, F. J. Ortega, and C. G. Vallejo, "
long autonomy or long delay?'the importance of domain in opinion mining," *Expert Systems with Applications*, vol. 40, no. 8, pp. 3174–3184, 2013.

[53] M. d. P. Salas-Zárate, J. Medina-Moreira, K. Lagos-Ortiz, H. Luna-Aveiga, M. A. Rodriguez-Garcia, and R. Valencia-Garcia, "Sentiment analysis on tweets about diabetes: An aspect-level approach," *Computational and mathematical methods in medicine*, vol. 2017, 2017.

[54] C. Ma, A. Osherenko, H. Prendinger, and M. Ishizuka, "A chat system based on emotion estimation from text and embodied conversational messengers," in *Proceedings of the 2005 International Conference on Active Media Technology, 2005.(AMT 2005).*, IEEE, 2005, pp. 546–548.

[55] S. Aman and S. Szpakowicz, "Identifying expressions of emotion in text," in *International Conference on Text, Speech and Dialogue*, Springer, 2007, pp. 196–205.

[56] M. Yang, D. Zhu, and K.-P. Chow, "A topic model for building fine-grained domain-specific emotion lexicon," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014, pp. 421–426.

[57] S. A. Golder and M. W. Macy, "Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures," *Science*, vol. 333, no. 6051, pp. 1878–1881, 2011.

[58] J. Xu, R. Xu, Y. Zheng, Q. Lu, K.-F. Wong, and X. Wang, "Chinese emotion lexicon developing via multi-lingual lexical resources integration," in *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, 2013, pp. 174–182.

[59] Y. Rao, Q. Li, L. Wenyin, Q. Wu, and X. Quan, "Affective topic model for social emotion detection," *Neural Networks*, vol. 58, pp. 29–37, 2014.

[60] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*, 2008, pp. 1556–1560.

[61] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Affect analysis model: Novel rule-based approach to affect sensing from text," *Natural Language Engineering*, vol. 17, no. 1, pp. 95–135, 2011.

[62] J. Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," in *Proceedings of the ACL student research workshop*, 2005, pp. 43–48.

[63] X. Ouyang, P. Zhou, C. H. Li, and L. Liu, "Sentiment analysis using convolutional neural network," in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, IEEE, 2015, pp. 2359–2364.

[64] L. Buitinck, J. Van Amerongen, E. Tan, and M. de Rijke, "Multi-emotion detection in user-generated reviews," in *European Conference on Information Retrieval*, Springer, 2015, pp. 43–48.

[65] D. Zhou, X. Zhang, Y. Zhou, Q. Zhao, and X. Geng, "Emotion distribution learning from texts," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 638–647.

[66] H. Yang and Y. Li, "Identifying user needs from social media," *IBM Research Division, San Jose*, p. 11, 2013.

[67] B. Liu, M. Hu, and J. Cheng, "Opinion observer: Analyzing and comparing opinions on the web," in *Proceedings of the 14th international conference on World Wide Web*, 2005, pp. 342–351.

[68] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Opinionfinder: A system for subjectivity analysis," in *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, 2005, pp. 34–35.

[69] C. Xian-Yi11, Z. Ling-ling, Z. Qian, and W. Jin, "The framework of network public opinion monitoring and analyzing system based on semantic content identification," *Journal of Convergence Information Technology*, vol. 5, no. 10, pp. 1–5, 2010.

[70] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle," in *Proceedings of the ACL 2012 system demonstrations*, Association for Computational Linguistics, 2012, pp. 115–120.

[71] Z. Li, S. Zhu, H. Hong, Y. Li, and A. El Saddik, "City digital pulse: A cloud based heterogeneous data analysis platform," *Multimedia Tools and Applications*, vol. 76, no. 8, pp. 10 893–10 916, 2017.

[72] D. McDuff, K. Rowan, P. Choudhury, J. Wolk, T. Pham, and M. Czerwinski, "A multimodal emotion sensing platform for building emotion-aware applications," *arXiv preprint arXiv:1903.12133*, 2019.

[73] A. Nijholt, "Smart, affective, and playable cities," in *Interactivity, Game Creation, Design, Learning, and Innovation*, Springer, 2018, pp. 163–168.

[74] B. Guthier, R. Alharthi, R. Abaalkhail, and A. El Saddik, "Detection and visualization of emotions in an affect-aware city," in *Proceedings of the 1st International Workshop on Emerging Multimedia Applications and Services for Smart Cities*, 2014, pp. 23–28.

[75] A. Landowska, "Affect-awareness framework for intelligent tutoring systems," in *2013 6th International Conference on Human System Interactions (HSI)*, IEEE, 2013, pp. 540–547.

[76] N. Thakur and C. Y. Han, "A complex activity based emotion recognition algorithm for affect aware systems," in *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, IEEE, 2018, pp. 748–753.

[77] T. H. M. De Oliveira and M. Painho, "Emotion & stress mapping: Assembling an ambient geographic information-based methodology in order to understand smart cities," in *2015 10th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, 2015, pp. 1–4.

[78] R. Alharthi, B. Guthier, and A. El Saddik, "Recognizing human needs during critical events using machine learning powered psychology-based framework," *IEEE Access*, vol. 6, pp. 58 737–58 753, 2018.

[79] G.-h. Xu, J.-x. Zhao, H.-y. Yang, and S. Liu, "A review of text feature extraction methods," *Software Guide*, vol. 17, no. 5, pp. 13–18, 2018.

[80] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[81] G. R. Semin and K. Fiedler, "The linguistic category model, its bases, applications and range," *European review of social psychology*, vol. 2, no. 1, pp. 1–30, 1991.

[82] B. Liu, S. Tang, X. Sun, Q. Chen, J. Cao, J. Luo, and S. Zhao, "Context-aware social media user sentiment analysis," *Tsinghua Science and Technology*, vol. 25, no. 4, pp. 528–541, 2020.

[83] R. Alharthi and A. El Saddik, "A multi-layered psychological-based reference model for citizen need assessment using ai-powered models," *SN Computer Science*, vol. 1, no. 5, pp. 1–31, 2020.

[84] K. Chodorow, *MongoDB: the definitive guide: powerful and scalable data storage*. " O'Reilly Media, Inc.", 2013.

[85] J. H. Allen, S. Barnum, R. J. Ellison, G. McGraw, and N. R. Mead, *Software security engineering*. Pearson India, 2008.

[86] R. M. Redlich and M. A. Nemzow, *Data security system and method*, US Patent 7,103,915, 2006.