A Depression Diagnostic System using Lexicon-based Text Sentiment Analysis

Bernice Yeow Ziwei, Hui Na Chua
Department of Computing and Information Systems, Sunway University, Selangor, Malaysia.
Berniceby@outlook.com, huinac@sunway.edu.my

Abstract— Clinical psychologists typically diagnose depression via a face-to-face session, applying depression diagnostic criteria. However, past literature revealed that most patients would not seek help from doctors at the early stage of depression, resulting in a declination of their mental health condition. Many people feel more comfortable sharing their thoughts online through social media platforms in today's modern digital era. Since then, many researchers have studied using social media to predict mental health conditions. To the extent of our knowledge, there is no study related to the experimentation of online depression diagnostic systems using text from social media platforms available for individuals. Our study presented in this paper has two-fold: i) enhancing existing lexicon-based methods by formulating a more accurate classification function for detecting depressive text from a social media platform, and ii) developing a depression diagnostic system embedded with our improved lexicon method for individuals to visualize their depression state instantly via an online interface. The depression lexicon developed in this study was validated by psychologists who have relevant domain knowledge in depression. Our experimented lexicon-based method achieved a precision of 77% and an F1-score of 74% in classifying depression state. In addition, we also found that depressed person uses more offensive words and are more aggressive when they communicate.

Keywords— Keyword Extraction; Lexicon-based sentiment analysis; Depression classification; Social media analytics; Text mining

I. Introduction

Depression is one of the significant social issues, and it is increasing daily. Millions of people suffer from depression, and only a tiny fraction of them undergo proper treatment. According to World Health and Substance Use [1], an approximation of 300 million individuals around the world endures depression. These numbers continue to increase with the rise in world population, and there is a lack of global provision for diagnosing, supporting, and controlling depression [1].

In the past, clinical psychologists diagnosed depression via a face-to-face session, applying the diagnostic criteria written by a professional psychologist [2]. However, the growing popularity of social media sites has introduced many researchers to a contemporary platform for diagnosing depression. This new platform eases the difficulties associated with screening and diagnosis of depression experienced by many psychologists in the past [3 – 4]. In addition, the growing use of social media strongly increases the possibility of utilizing the internet as a medium to discover the world and express one's opinions [5].

Numerous psychologists are turning their sights to online media. The correlation between the depressive state and their tweet sentiment was emphasized by [6]. In addition, Cavazos-Rehg et al. [6] affirm that symptoms of major depressive disorder can be traced in social media posts after succeeding several analyses on contents written by depressed individuals. However, most of the prior studies' models are not available used by individuals. This limitation restricts individuals from running depression diagnostic analysis to inform their depression state instantly via an online interface. Therefore, this research aims to fill this gap by developing a real-time depression diagnostic system with this capability. On top of that, we examine the performance of several depression lexicons constructed by prior researchers and found that more research efforts are still needed to enhance those lexicon terms to classify depressive text more accurately. Therefore, this research works on enhancing various lexicon [7-9] through expanding the depression diagnostic criteria and extending our preliminary work in [10] to derive a more accurate classification function for classifying depression states.

Our research posits that there are similarities between the mental state of individuals and the sentiment of their tweets. This research comprises text analysis that focuses on drawing insights from written communication to determine whether a tweet is related to depressive thoughts. The system used in this research is voluntary because an individual is fully responsible for acquiring access

[1] World Health and Substance Use
https://www.who.int/teams/mental-health-and-substance-use/suicide-data

to and actively participating in the screening process. Individuals keen on knowing their depression state can voluntarily utilize our system.

At this point, we must add important adequacy. First, in this research, we do not diagnose real depressed patients but only on textual data written by them through social media. Therefore, we restrict ourselves only to the modest objective of identifying depressive-indicative social media posts by using depression-associated terms in the text, i.e., a lexicon-based sentiment analysis approach.

II. Related work

A. *Sentiment Analysis in Social Media*

The fundamental symptoms of a depressed person are acute negative emotions and the absence of positive emotions. One of the suitable techniques for depression detection is sentiment analysis. It aims to discover the sentiment polarity of text that belongs to a particular person and uncover the user's opinions [15]. Sentiment analysis on social media, such as Twitter (a microblogging application) data involves the extraction of emotions and opinions from Twitter posts (tweets) [16]. Twitter users would share their feelings and opinions over various topics such as retirement, investment, health, and politics [16]. Hence, it can be the source of data and information in this research.

Sentiment analysis is very beneficial when it comes to analyzing the behaviors and feelings of an individual [17]. Work in [18] and [19] justified that twitter is suitable for sentiment analysis, those researchers have demonstrated early and existing results on sentiment analysis using Twitter data. Furthermore, the idea of using text analysis to extract sentiments related to depression was supported by [20].

B. *Lexicon-based Text Sentiment Analysis*

One common way to conduct sentiment analysis is by matching each word in a text with words in a dictionary that contains semantic orientation and deriving a sentiment score for the text [41]; this is known as the lexicon-based approach for sentiment analysis. Similarly, Almatarneh and Gamallo [28] explained that the lexicon-based approach employs a sentiment lexicon to derive polarity value for each text following a basic mathematical algorithm.

Taboada et al. [41] introduced a word-based method to obtain text sentiment. The author extends the SO-CAL system [42] by including other parts of speech categories, for instance, adding additional adjectives, nouns, verbs, adverbs, intensification, negation, irrealis blocking, and many more. The extension leads to an increase of up to 70% accuracy and a statistically significant improvement from previous instantiations of the SO-CAL system ($p < 0.05$). Moreover, the researcher demonstrated that manually built corpus provides a solid foundation for a lexicon-based approach. Finally, the author concluded that lexicon-based methods for sentiment analysis gave a proven cross-domain performance and can be easily extended to multiple sources of knowledge.

Zhang et al. [43] adopted a lexicon-based approach to perform entity-level sentiment analysis on tweets. The researcher first adopted the lexicon-based approach for entity-sentiment analysis, then improved the overall result by including additionally identified tweets based on the information derived from the result of the lexicon-based method. Next, the classifier assigned sentiment polarities to the words in the newly identified tweets. This assignment significantly improved accuracy, precision, recall, and F1-Score, outperforming state-of-the-art baselines. Furthermore, compared to Taboada et al. [41], researchers from [43] label the words automatically instead of manually annotating the words in the lexicon.

Palanisamy et al. [44] created a sentiment lexicon built from Serendio taxonomy [45]: positive, negation, negative, and stop words. After constructing the lexicon, the researcher experimented with several preprocess methods. The best preprocess method selected was stemming, emoticon detection, exaggerated words, normalisation, hashtag detection, and shortening words. After that, the researchers built a lexicon-based system to classify tweets into positive and negative sentiment based on the contextual sentiment orientation of words in tweets. The best result obtained was an F1-score of 0.8004.

C. *Factors and Measurements for Classifying Depression*

Many works on depression were carried out in medicine and psychology to explain the factors that cause depression. Hence, numerous measures, scales, and criteria had been established to diagnose depression:

- Primary Care Evaluation of Mental Disorders was reduced into three-page patient health questionnaires for nine depression symptoms [11 – 12]. It also encompasses common psychiatric issues: depression, alcohol, somatoform, eating disorders, and anxiety. It provides a severity score for each symptom, and the scores can be utilized to follow outcomes.
- Center for Epidemiological Studies Depression (CES-D) Scale [13] comprises 20 questions anent the mental state of a person. User has the choice to answer the questions regarding the severity of their condition. The mental state can be determined based on a scale of the total score. The questionnaire contains symptoms of depression ranging from a feeling of guilt and worthlessness, depressed mood, feeling helpless and hopeless, loss of appetite, sleep disturbance, and psychomotor.
- Diagnostic and Statistical Manual of Mental Disorders (DSM) [14] consists of nine types of depression indicators. Psychiatrists and clinicians commonly use it to examine depression symptoms during a specified duration of time. These criteria have been well validated and utilized

in numerous real-world cases for a long time. It consists of symptoms, descriptions, and other criteria for diagnosing mental illness.

In practice, DSM-IV took 12 years to evolve into DSM-V. It demonstrates that psychologists will need a longer timespan to upgrade a psychological measure to the latest version. However, psychological metrics and measurements may not comprehensively scrutinize new behavioural factors. As an example, symptoms and behaviours are manifest on social media. In the context of all these challenges, this research aims to investigate the likelihood of using social media data to identify depressive thoughts.

D. *Past studies on depression analysis*

Researchers have been working on depression studies using a supervised machine learning approach. Park et al. [21] concluded that a correlation exists between the depressive state and its tweet sentiment. De Choudhury et al. [22] developed a statistical metric: Social Media Depression Index to complement this idea further. Malmasi et al. [23] developed a colour-coded-basis algorithm that determines the severity of depression in Twitter posts together with SVM – RBF and Random Forest classifier. Moreover, Ay et al. [24] employed convolutional neural networks and long-short term memory architectures to detect depression using electroencephalogram signals.

Alternatively, a researcher employed affective computing technology to research depression. Zucco et al. [25] adopted sentiment analysis and affective computing methods for depression detection. According to the research, sentiment analysis was used to derive the sentiment polarity. In contrast, affective computing was employed to detect emotion in text and an individual's affective state, including a person's physical aspect, i.e., speech and gestures and physiological signals, respiration, and heart rate. Lastly, the researchers developed a mobile application that collects users' data unobtrusively for depression conditions monitoring.

The linguistic features of a depressed individual were researched by Ramirez-Esparza et al. [26] to investigate the linguistic markers of depression through acquiring posts of depressed and non-depressed individuals from Internet forums. It examines the text with LIWC, a word counting software. The result reveals that the online depressed users use more negative emotion words but less positive emotion words, more first-person singular pronouns, and fewer first-person plural pronouns. In addition, Cheng et al. [27] concluded that words related to lowered self-esteem and guilt are commonly found in text posted by depressed people after reviewing several questionnaires for depression diagnostics like ICD-10 and DSM-5 and conducting interviews with several groups of mental health professionals.

While there is a vast amount of work on depression analysis and depression lexicon development, Losada et al. [8] examined and enhanced lexical resources for detecting symptoms of depression. The research suggested two strategies for lexicon enhancement: thesaurus-based and corpus-based. Corpus-based methods rely on distributional similarity generated by distributional semantic models, containing representations of explicit context or embeddings trained on large corpora. The Thesaurus approach uses lexical relationships found in lexical resources, such as synonyms, to construct domain-specific corpus.

E. *Related Studies on Depression Classification*

Prior works [21 – 24] employ deep learning, machine learning, and statistical methods for classifying depression. Alternatively, this study explores a lexicon-based approach to classify depressive-indicative social media posts. We must emphasize that this research focused on creating a depression lexicon and discovering the best formula for the depression lexicon. Future research can investigate whether adding the depression lexicon terms as a feature used in machine learning models would give a better classification result and suggest or develop machine learning models that work best with the depression lexicon for depression detection.

Taboada et al. [29] concluded that the lexicon-based method is one of the robust techniques for sentiment analysis; it can be improved easily without needing to train all over again. Moreover, the researcher demonstrated that the lexicon method had proven cross-domain performance. In the subsequent paragraphs, we will elaborate on prior studies that conducted lexicon-based methods for depression analysis.

Wang et al. [30] adopted Chinese vocabulary from HowNet [31] to create a Chinese lexicon. Several Chinese artificial rules were formulated to comprehend the linguistic rule of Chinese text; eventually, this artificial rule is used to derive the depression inclination of each microblog text. Subsequently, ten features related to the user's interaction, the behaviour of a user, and the content of microblog posts of depressed persons were acquired to create several supervised models for classification. After that, microblogs[2] were extracted for a group of psychologists to classify all the blogs into depressed or standard categories. Waikato Environment for Knowledge Analysis was utilized to classify users into depressed or normal categories. Furthermore, the researcher built several supervised models to obtain a more reliable result. The final result of 80% precision was obtained with 10-fold cross-validation on the best model (Bayes Network).

[2] Weibo https://weibo.com/login.php

```
For each word in the text do
    if the word is in the positive list then
        sum = sum +1
    else
        if the word is in the negative list then
            sum = sum -1
        End if
    End if
End for
```

Eq (1) Pseudo-algorithm to label each word

$$score = \frac{\Sigma \ \ sum}{n}$$

Eq (2) Formula to generate a score for each sentence

```
If score >= 0.2 then
    label = 1
else
    if -0.5 < score < 0.2 then
        label = 0
    else
        if score <= -0.5 then
            label = -1
        End if
    End if
End if
```

Eq (3) Pseudo-algorithm to label each sentence

Basantani et al. [7] adopted a depression lexicon that was freely available in github.com and constructed formulas to annotate depressive text: Eq (1), Eq (2), and Eq (3).

In reference to Eq (1), the method begins by looping through all words in a text; if a word is detected as negative, the algorithm will deduct 1 from the sum. On the contrary, if a positive word is detected, the algorithm will add 1 to the sum as neutral. Eq (2) generates a score by taking the total sum from Eq (1) divided by the total number of words in the text.

Eq (3) classify text based on the score obtained from Eq (2); if the value is more than or equal to 0.2, the text will be labelled as not depressed. Otherwise, if the value is less than 0.2 and more than -0.5, the text will be classified as neutral, or else if the value is less than or equal to -0.5, it is labelled as depressive text. Subsequently, Naive Bayesian, Decision Tree, Support Vector Machine Classifier, K Nearest Neighbor, and Random Forest classifiers were trained on the labelled dataset for predicting depressive text.

Losada and Gamallo [8] used two methods to enhance depression lexicons [9, 32]. One of the methods adds vocabularies from Wikipedia by using distributional similarity to select the most similar words in lexicons constructed. Another method utilizes Wordnet-based expansion, which selects synsets similar to each lemma of the original lexicon [9, 32] to eliminate dissimilar and unrelated words after expanding vocabularies. The proposed solution aims at expanding non-ambiguous words that only contain one synset in Wordnet.

Subsequently, the researchers utilized a document corpus from [33] consisting of 500 user submissions from depressed and non-depressed individuals. Following that, all posts were extracted for every user listed in the document corpus, making up an extensive sequence of submissions for each user. This document was then separated into two parts, DLU16A, and DLU16B, which are utilized to evaluate the enhanced lexicon. Finally, the author carried out a document search task to evaluate the enhanced lexicon, which involves searching through the documents and identifying individuals at risk of depression. Three models were created for ranking:

1) Ranks each individual randomly
2) Four measures were used to determine whether the user is at risk of depression: sad, hopeless, loneliness, and worthlessness
3) The enhanced lexicon was used to detect signs of depression in each user submission.

The result revealed that the enhanced lexicon performs the best among all three ranking models because expanding the dictionary incorporates more relevant terms to detect signs of depression.

Neuman et al. [9] put forward a systematic procedure to produce a depression lexicon. The process starts with an automated system that gathers resources from Bing. Subsequently, Stanford parser was adopted to produce dependency representation for each sentence found in those web pages. After that, a system for constructing lexicon – Pedesis [9] was invented to create a depression lexicon. Subsequently, depression measurements – DepScale [9] were formulated with the depression lexicon to classify depressive-indicative text. DepScale was based on the sum of several variables:

1) Percentage of phrases detect from the lexicon over the total number of phrases in the text
2) Percentage of different phrases detected from the lexicon over the total number of phrases in the text
3) Percentage of first-person pronoun words occurred in the text.

This method was later tested on questionnaires from Mentalhelp.net , blog authorship corpus [32] and validated through expert judgment

There are several issues and limitations seen in literature [7 – 9]; researchers [7] and [8] did not validate their methods and lexicon using expert judgment. As a result, the researcher [8] pointed out that the researchers had experienced a lack of proper annotated resources to validate the performance of their lexicon on search tasks.

Moreover, the researchers [9] have not tested real-life cases like social media text. Their depression lexicon was developed using a bottom-up method that obtains opinions and text from a depressed person. In contrast, our study incorporates judgment and opinions from depressed patients and mental health experts and tests on actual social media text. In this study, we examine and enhance the performance of depression lexicons created by [7 – 9]. We share a similar research scenario of using a lexicon-based method on English text that involves creating a depression lexicon and defining several measurements or formulas for classifying depressive text. As a result, we also improved the lexicon and methods used in our previous research [10].

III. RESEARCH METHODOLOGY AND SYSTEM FRAMEWORK

A. *Data Collection*

The first step involves collecting depressing and non-depressing tweets from the Twitter platform. The sample size was 2890 tweets consists of 1445 depressing tweets and 1445 non-depressed tweets. The depressed tweets were collected following the keywords defined by [7], and non-depressed tweets were acquired, excluding the keyword defined by [7].

In addition, we also obtained lexicons from [7 – 9]. Hate speech and offensive language lexicon [35] were adopted in this research as it was evinced by [36] that a depressed person will become more aggressive when communicating with others. In the subsequent section ‘c’, all words from lexicons [7 – 9, 35] will be added to the lexicon [10], and in section 'D', the synonyms for words in those lexicons were generated.

B. *Thesaurus-based method*

For our depression lexicon, the synonyms were generated through thesaurus.com. Notably, not all synonyms generated by thesaurus.com are akin to the semantic meaning of the searched word. For instance, the word "down" and "blue" does not necessarily have the same semantic meaning as the word "sad". Therefore, this research aims to extract the terms indicated as "most relevant" by thesaurus.com and discard the rest of the irrelevant words.

C. *Expanding the depression lexicon using Keyword Extraction*

We acquire a similar data source described by [37] to construct a depression lexicon for this method. Rapid Automatic Keyword Extraction [38] was utilized to extract keywords from the following data sources. Subsequently, keywords extracted will be input into the lexicon and evaluated using expert judgment in section ‘D’.

1) *Professional Twitter Accounts:* We acquire data from Professional mental health Twitter accounts nominated by The Breakthrough Voice[3]. The reason behind collecting tweets from professional mental health Twitter accounts is that they are an academic community with deep knowledge of mental health.

2) *Depression Blogs:* In addition to collecting data from Twitter, an extra source of information can be attained from web blogs[4] related to depression. For each listed website, the scraping will commence on the webpage that displays a library of depression blogs and extend to deeper links where it includes all depression weblogs written on the website. Most of the blogs include daily scenarios and diaries that explain depression and what the person felt about depression. Having an additional platform to scrape information about depression is extremely useful for building a depression lexicon.

D. *Rating words in the depression lexicon*

After adding keywords to enhance the depression lexicon, we validate the depression lexicon. We recruited psychologists and depressed people as participants in our study that have proven domain knowledge on depression. Participants were requested to rate each word in the lexicon following the scale design. The scale is defined in Table 1:

Initially, we have sent invitations to 6 psychologist graduates that have previously studied depression and four depressed people to participate in the rating session. Those participants had undergone a pre-test survey to examine their knowledge of depression. During the pre-test survey, we tested participants on depression diagnostics and the content listed on depression diagnostic screening tools. The survey was distributed online via a google survey form.

After filtering those potential candidates, we recruited two depression experts (psychologist graduates) and two depressed patients to rate the words in the depression lexicon. After all of the participants rated the words in the lexicon, we classified each word in the lexicon. Previous literature concluded that the average and means score might give misleading results [39]. To tackle the potential bias, we classify words in the lexicon based on the score most of the raters agreed on. The categories of words in the lexicon are defined as follow:

- Positive words typically imply positive sentiment, such as kindness, happiness, and satisfaction (rated 1 or 2 by more than half of the raters)
- Negative words consist of words linked to depression (rated 4 or 5 by more than half of the raters)

[3] The Breakthrough Voice https://inbreakthrough.org/twitter-mental-health-self-care/

[4] Top Stories http://www.healthline.com/health/depression/best-blogs-of-the-year

- Definite Negative words are words rated as related to depression by all raters.
- Neutral words do not have positive sentiment and are not linked to depression (rated 3 by more than half of the raters).

Subsequently, in section 'E' we will evaluate several preprocessing methods and determine the most suitable preprocessing method to be used in this research. Following that, Section 'F' will explain the experimentation settings used to derive the best classification function. Finally, formulas used for evaluation are listed in section 'G'.

TABLE I
RATING SCALE

Rating Score	Meaning
1	Expresses a strong positive meaning
2	Indicates a positive meaning or has the opposite meaning of depression
3	Having a neutral state does not imply positive sentiment and is not related to depression
4	Related to depression
5	Very related to depression

E. *Data Pre-processing*

In this section, we constructed several preprocessing methods and conducted several evaluations to determine the most suitable preprocessing used in the research. We evaluate the preprocessing methods by comparing the classification function's performance with the raters' results. A suitable preprocessing method would improve the result of the classification function. Subsequent section ' F ' will describe the rating session and evaluation steps. The preprocessing methods used and the results obtained are shown in Table 2.

F. *Experimentation setting to derive the best classification function*

After finalizing the words in the depression lexicon, we got lists of positive, negative, and definite negative words. We then defined and tested several formulas using lists of positive, neutral, negative, and definite negative words for depression detection. All of the formulas are defined in Table 3. All of the formulas are targeted to detect whether the text is depressive-indicative or not depressive-indicative.

The sample size of 2890 tweets consists of depressing tweets collected following the keywords defined by [7] and non-depressed tweets acquired without the keyword defined by [7]. Although the researcher [7] defined keywords to gather depressing tweets, we decided to have a rating session to include human judgment in verifying whether the text is depressed or not depressed. Therefore, we decided to have another rating session; seven participants were employed to rate a sample size of 2890 tweets. The rating team consists of one psychologist, three psychologist graduates, one depressed person, and two computer science graduates. Those raters were tasked to rate the 2890 tweets on whether each of those tweets is depressive or non-depressive indicative.

After ensuring all text is appropriately annotated, we compare the mode of those 7 participants' rating scores with the results generated by each prepossessing method in Table 2 and several classification functions constructed in Table 3.

In addition, we adopted the depression lexicon and methodology from [9] that uses DepScale to determine whether depression is the topic of a text. DepScale has three metrics: a ratio of depressed words found over the total number of words in a text, ratio of different words from their depression lexicon of the total number of words in the text, and percentage of first-person pronoun words that occurred in the text. Furthermore, we adopted the lexicon constructed by [7] and followed the methodology created by [7], which classified depressive text using Eq(1), Eq(2), and Eq(3). Lastly, we utilised the lexicon created by [8] to compare the result with our proposed classification function.

G. *Evaluation to derive the best formula*

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$$

Eq (3) Precision

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative}$$

Eq (3) Recall

$$F1 - Score = \frac{(2 * Precision * Recall)}{(Precision + Recall)}$$

Eq (3) F1 Score

Recall, precision, and F1 score are metrics used to evaluate the classification function on classifying depressive text. The metrics were used as the primary indicator because the rating scale for the survey was in categorical value: depressed and not depressed.

In reference to the above formulas, positive refers to the depressive text, and negative refers to the non-depressive text. True positive refers to records that are predicted as positive and are actual positive records. On the contrary, false-positive refers to records that are predicted as positive records but are negative. Finally, false-negative refers to predicted as negative but are positive records.

To obtain a precision score, we use Eq (4), the equation computes the number of records that are predicted correctly as positive among all the records that are predicted as positive. On the contrary, recall can be obtained using Eq (5), the formula computes the number of records positive records that are predicted correctly as positive among all positive records. Lastly, the F1 score is obtained using Eq (6), the harmonic average between recall and precision. These formulas are commonly used to evaluate classification and prediction tasks [40].

We evaluate the formula based on the F1 score, the average weight score of precision and recall; it incorporates precision and recall equally important. Precision measurement tells us that among all the text classified as depressed, how many of those records are depressed, and recall tells us out of actual depressive text how many of those texts is classified correctly as depressive.

It is also worth noting that it is essential to consider both recall and precision because it enables us to look into a bigger picture of how the classification function performs on classifying depressive text. Using the F1 metric includes both precision and recall that incorporate false negative, false positive, and true positive records for evaluation. Those three measures are related to depressed records and should be included for evaluation.

In addition to precision, recall, and F1 score, we also use Cronbach Alpha to measure the degree to which results obtained from the best classification function and rating score from raters' judgment are homogenous.

In the subsequent section, 'H', we will explain the flow and processes of the depression diagnostic system.

H. *SYSTEM ARCHITECTURE AND ITS FUNCTIONAL COMPONENTS*

The proposed system architecture is diagrammatically depicted in Fig. 1.

The web application user interface acquires the user's inputs, i.e., Twitter username and choice of activity that the user wants the web application to perform. The user input will be transferred to the system's backend (web application server). The backend system will then signal the crawling data function to commence scraping the user's tweets from the Twitter data source using Twitter API. After acquiring data from the Twitter data source, the data crawling function will store those tweets in a data mart.

Next, those tweets will be inputted into the data transformation function to preprocess those tweets. Subsequently, the transformed tweets will be sent back to the data mart and loaded into the classification function to conduct depression analysis on tweets. Finally, the result of depression analysis will be transmitted into the web application user interface via main server logic (web application server). Subsequently, the user interface will utilize various visualization features to display the results of depression analysis.

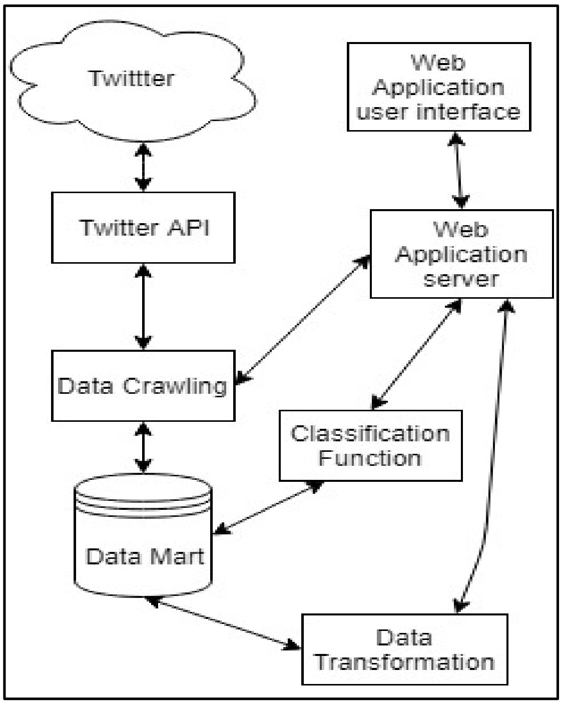


Fig. 1 The proposed application system architecture.

IV. RESULTS

A. *Experimentation result*

TABLE IIIII

RESULTS FOR DIFFERENT PREPROCESSING METHODS.

No.	Preprocessing method	Precision (%)	Recall (%)	F1-Score (%)
1	Convert text to lower case, convert English contraction to a complete form, remove non-alphabetical words, replace three or more consecutive letters into two consecutive letters	77.43	70.66	73.89
2	Convert text to lower case, convert English contraction to a complete form, remove non-alphabetical words, replace three or more consecutive letters into two consecutive letters, remove URL links	75.95	69.48	72.57
3	Convert text to lower case, convert English contraction to a complete form, remove non-alphabetical words, replace three or more consecutive letters into	75.75	69.38	72.42

	two consecutive letters, remove URL links, remove @ mentions			
4	Convert text to lower case, convert English contraction to a complete form, remove non-alphabetical words, replace three or more consecutive letters into two consecutive letters, remove URL links, remove @ mentions, remove hashtags	76.68	69.16	72.73
5	Convert text to lower case, convert English contraction to a complete form, remove non-alphabetical words, replace three or more consecutive letters into two consecutive letters, remove URL links, remove @ mentions, remove hashtags, remove stop words	77.62	69.76	73.48
6	Convert text to lower case, convert English contraction to a complete form, remove non-alphabetical words, replace three or more consecutive letters into two consecutive letters, remove URL links, remove @ mentions, remove hashtags, remove stop words, correct spelling of words	64.63	42.99	51.63

We select the best preprocessing method based on the highest f1 score; therefore, preprocessing method 1 is chosen as the preprocessing method used in this research. Based on the results in Table 2, we noticed that correcting the spelling of words resulted in poor performance because misspelled words and typos have numerous variations, making it complex to correct the spelling of words accurately. In addition, the URL links, mentions, and hashtags are essential features to be included for depression analysis because URL links like a suicide hotline, retweeting or mentioning mental health professional accounts, and depression hashtags are relevant indicators of depressive text. Furthermore, removing stop words does not increase the performance of the classification function because it consists of the first-person pronoun like I, mine, me, my, which a depressed person commonly uses; this finding was consistent with [9].

TABLE IVVVI

RESULTS OF SEVERAL CLASSIFICATION FUNCTIONS AND PRIOR WORK.

No.	Formula	Precision (%)	Recall (%)	F1-Score (%)
1	$\Sigma\ DefiniteNegative > 0$ OR $(\Sigma Negative > \Sigma Positive)$ (this research)	77.43	70.66	73.89
2	$\Sigma\ DefiniteNegative > 0$ OR $((\Sigma Negative > \Sigma Positive)\ AND\ (\Sigma Negative > \Sigma Neutral))$ (this research)	71.43	58.15	64.11
3	$\frac{positive - negative}{positive + negative}$ (this research)	39.78	48.80	43.83
4	$\frac{positive - negative}{positive + neutral + negative}$ (this research)	39.78	48.12	43.56
5	$\frac{Positive + neutral - negative}{positive + neutral + negative}$ [10]	49.15	76.67	59.90
6	Neuman et al. [9]	53.75	100	69.92
7	Basantani et al. [7]	78.95	33.33	46.88
8	Losada and Gamallo [8]	53.17	76.65	62.79
9	Remove words listed in hate speech and offensive language created by [35]	74.74	70.62	72.62

In reference to Table 3, formulas 1 to 4 are formulas formulated by this research, and formulas 5 to 9 are from prior work. Formula 3 to 5 classifies a text as depressed if the score obtained is a negative value; otherwise, the text will be classified as not depressed. Formula 9 follows the first formula but excluding the words from hate speech and offensive language [35]

More efforts are needed to enhance the performance of the depression lexicon and methodology done by [7 - 9], as shown in Table 3. In our research, we increased the F1 score by approximately 4% compared to the previous work done by [9]. Additionally, we increased the F1 score to approximately 27% compared to the result obtained from prior work [7]. Furthermore, the sixth formula achieved 100% recall because the result shows that out of 2764 texts, there were 2719 texts classified as depressed. In this case, the recall will be 100% because all actual depressive texts are classified as depressed, and none of the actual depressive texts are classified as not depressed since there are 2764 texts already predicted as depressed. However, 1253 actual non-depressive text is classified as depressed in precision. As a result, this method performs poorly in terms of precision. Therefore, we used the F1 score to validate and examine the classification function to avoid misinterpreting the result. The best formula selected is the first formula since the first formula gave the highest F1 score; therefore, the first formula is the proposed classification function in this research.

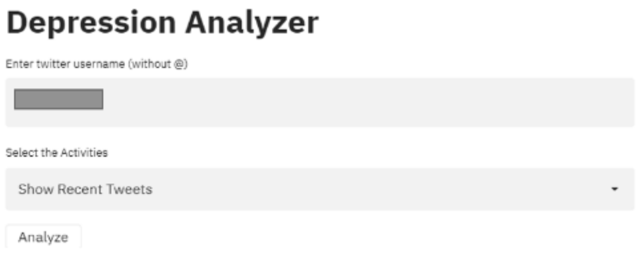
The proposed classification function increased the F1 score to approximately 11% compared to the result generated from the lexicon constructed by [8]. Lastly, we noticed that adding words from hate speech and offensive language created by [35] increased the performance of the classification function by approximately 2% in terms of F1 score. This finding shows that many depressed people will use offensive words and are more aggressive when communicating with people; this insight is consistent with the findings from [36]. In comparison to our previous research in [10], we managed to increase the precision of the classification function by approximately 28%.

Furthermore, we achieved a Cronbach Alpha score of 0.750 with a Precision of 77.43%. These results indicate an acceptable level of internal consistency and accuracy. Therefore, our proposed Depression values are "Depressed" and "Not Depressed". Compared to our previous research [10], we only achieved a Cronbach Alpha score of 0.70 previously.

B. *Proposed web application*

This section presents the outputs and visualizations of the proposed web application.

Fig. 2 presents the scenario when a user is prompted to enter their Twitter username and select the activity to be performed. In reference to Fig. 3, under 'select the activities', the default option is 'show recent tweets'. Upon choosing this option, the web application will display a data frame for the user to view their tweets together with the number of depressive words, non-depressive words in each text, and the score column that shows the classification results obtained from the classification function.



Depression Analyzer

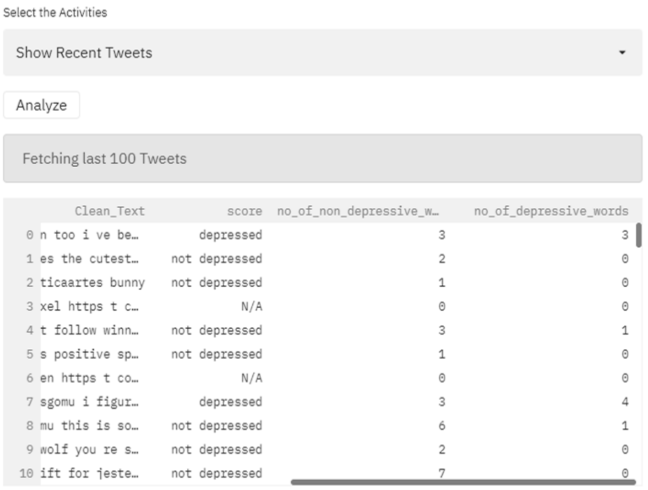
Enter twitter username (without @)

[redacted]

Select the Activities

Show Recent Tweets

Fig. 2 User Input.



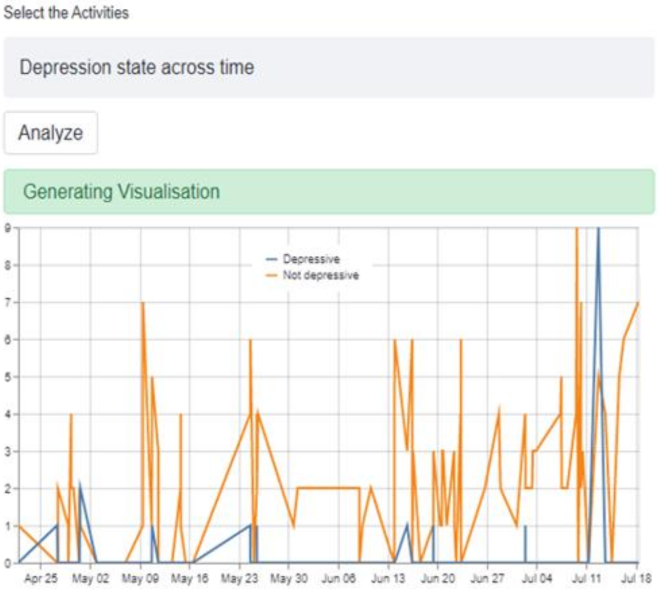
Select the Activities

Show Recent Tweets

Fetching last 100 Tweets

```
     Clean_Text        score          no_of_non_depressive_w…  no_of_depressive_words
 0 n too i ve be…      depressed                            3                       3
 1 es the cutest…      not depressed                        2                       0
 2 ticaartes bunny     not depressed                        1                       0
 3 xel https t c…      N/A                                  0                       0
 4 t follow winn…      not depressed                        3                       1
 5 s positive sp…      not depressed                        1                       0
 6 en https t co…      N/A                                  0                       0
 7 sgomu i figur…      depressed                            3                       4
 8 mu this is so…      not depressed                        6                       1
 9 wolf you re s…      not depressed                        2                       0
10 ift for jeste…      not depressed                        7                       0
```

Fig. 3 User’s Tweets



Select the Activities

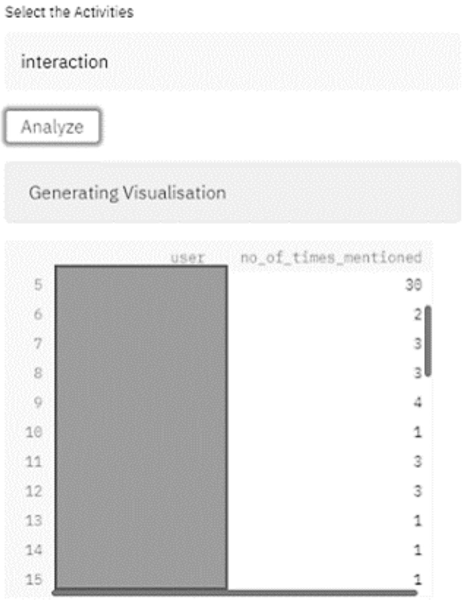
Depression state across time

Generating Visualisation

Fig. 4 Depressive state of a user across time

The line graph in Fig. 4 visualizes the trend of the depressed and non-depressed tweets of a particular depressed Twitter account. From the line graph, a Twitter user can discern whether they are having depressive thoughts at a timeframe and observe the changes of their thoughts across time. In this real-life example, the user was diagnosed with depression on the 13th of July 2021. Prior to that date, the user had been having several depressive thoughts. This trend was consistent with the result displayed on the line graph in Fig. 4; it showed a peak of depressive text on the 13th of July 2021, and several depressive texts occurred before the 13th of July 2021.

In addition, users of the web application can view interactions with other Twitter users on Twitter. For example, in Fig. 5, The table shows a list of users mentioned in the Twitter account and the number of times it was mentioned in the Twitter account. As we were using live data, we masked the users' information to protect their account privacy in Fig. 5.



Select the Activities

interaction

Generating Visualisation

	user	no_of_times_mentioned
5	[redacted]	30
6	[redacted]	2
7	[redacted]	3
8	[redacted]	3
9	[redacted]	4
10	[redacted]	1
11	[redacted]	3
12	[redacted]	3
13	[redacted]	1
14	[redacted]	1
15	[redacted]	1

Fig. 5 Interaction of Twitter account

V. CONCLUSION AND FUTURE WORK

A classification function that differentiates depressive-indicative text and non-indicative text based on lexicon-based method was proposed in this research. Overall, we built a depression lexicon and a classification function that achieved an F1 score of 74%, precision of 77%, and Cronbach Alpha of 0.750. The result was achieved by verifying a group of depression experts who validated wordings in the depression lexicon. Furthermore, the preprocessing phrase in this research was enhanced, which facilitates a better result in detecting depressive-indicative text.

The classification function is deployed on github.com, which researchers can use when they have limited resources or cannot find mental health professionals to annotate the depressive text. Finally, an online web application that consists of the proposed classification function was deployed online for the public to conduct depression analysis for free.

The first preprocessing method in Table 2 is the chosen method used to preprocess text before inputting it to the classification function. The first formula from Table 3 is the selected classification function used in this research. Compared to our previous research [10], we increased the precision of the classification function by approximately 28%, the F1 score rose to 14%, and Cronbach alpha of approximately 5%.

We compared our result with results generated by adopting prior work [7 – 9]. The F1 score increased up to approximately 27% compared to prior work [7], the F1 score went up approximately 11% compared to results generated from prior work [8], and the F1 score raised to 4% when compared to results obtained from prior work [9]. These results highlight limitations from prior work and suggest that additional works are needed to improve prior work's results. In this research, we have contributed to enhancing their classification method and prior work's lexicon.

Aside from that, there is an increase of 2% in F1 score when we add words from hate speech and abusive language [35]; this shows that depressed people will likely use more offensive words and are more aggressive when communicating with others. Lastly, we found that neutral words are not relevant in classifying depressive text.

The limitation of this research is the constraint set by Twitter API. At this stage, the analysis can only be performed on a limited number of tweets, i.e., 2,400, due to the restrictions enforced by Twitter. Furthermore, the depression diagnostic system can only leverage tweets set as 'public'. Future research can incorporate other social media platforms for depression analysis, such as Facebook and Instagram. Additionally, audio and images can be included for classifying depression. Apart from that, future research can investigate whether adding the depression lexicon terms as a feature in machine learning models would give a better prediction result and suggest or develop machine learning models that work best with the depression lexicon for depression detection.

VI. REFERENCES

[1] J. Dine, Companies, international trade, and human rights, Cambridge University Press, 2005.

[2] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, A depression detection model based on sentiment analysis in micro-blog social network. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 201-213. Berlin, Germany: Springer, 2013. https://doi.org/10.1007/978-3-642-40319-4_18

[3] S. Gilbody, T. Sheldon, and A. House, Screening and case-finding instruments for depression: a meta-analysis. Cmaj. 178(8), 997-1003, 2008. https://doi.org/10.1503/cmaj.070281

[4] A. J. Mitcehk and J. C.Coyne, Screening for Depression in Clinical Practice: An Evidence-Based Guide. OUP USA, 2009.

[5] C. Aggarwal, An introduction to social network data analytics. In Social network data analytics, pp. 1-15. Boston, MA: Springer, 2011. https://doi.org/10.1007/978-1-4419-8462-3_1

[6] P. A. Cavazos-Rehg, M. J. Krauss, S. Sowles, S. Connolly, C. Rosas, M. Bharadwaj, and L. J. Bierut, A content analysis of depression-related tweets. Computers in Human Behavior, 54, 351–357, 2016. https://doi.org/10.1016/j.chb.2015.08.023.

[7] A. Basantani, Y. Kesarwani, S. Bhatia, and S. Jain, EmoCure: Utilising Social Media Data and Smartphones to Predict and Cure depression. IOP Conference Series. Materials Science and Engineering, 1110(1), 12010, 2021. https://doi.org/10.1088/1757-899X/1110/1/012010

[8] D. E. Losada and P. Gamallo, Evaluating and improving lexical resources for detecting signs of depression in text. Language Resources and Evaluation, 54(1), 1–24, 2020. https://doi.org/10.1007/s10579-018-9423-1.

[9] Y. Neuman, Y. Cohen, D. Assaf, and G. Kedma, Proactive screening for depression through metaphorical and automatic text analysis. Artificial Intelligence in Medicine, 56(1), 19–25, 2012. https://doi.org/10.1016/j.artmed.2012.06.001

[10] B. Y. Ziwei, and H. N. Chua, An Application for Classifying Depression in Tweets. In Proceedings of the 2nd International Conference on Computing and Big Data, pp. 37–41, 2019.

[11] R. L. Spitzer, K. Kroenke, and J. B. Williams, Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. JAMA : the Journal of the American Medical Association, 282(18), 1737–1744, 1999.

[12] R. L. Spitzer, J. B. Williams, K. Kroenke, R. Hornyak, and J. McMurray, Validity and utility of the PRIME-MD Patient Health Questionnaire in assessment of 3000 obstetric-gynecologic patients: The PRIME-MD Patient Health Questionnaire Obstetrics-Gynecology Study. American Journal of Obstetrics and Gynecology, 183(3), 759–769, 2000. https://doi.org/10.1067/mob.2000.106580

[13] L. S. Radloff, The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. Applied Psychological Measurement. 1(3), 385–401, 1977.

[14] "Diagnostic and statistical manual of mental disorders," DSM-5, Fifth edition. American Psychiatric Publishing, 2013.

[15] B. Pang and L. Lee, Opinion mining and sentiment analysis. Comput. Linguist. 35.2, 311-312, 2009.

[16] A. Saxena, A Semantically Enhanced Approach to Identify Depression-Indicative Symptoms Using Twitter Data, 2018.

[17] S. Stieglitz and L. Dang-Xuan, Emotions and Information Diffusion in Social Media-Sentiment of Microblogs and Sharing Behavior. Journal of Management Information Systems, 29(4), 217–248, 2013. https://doi.org/10.2753/MIS0742-1222290408

[18] A. Go, R. Bhayani, and L. Huang, Twitter sentiment classification using distant supervision. CS224N project report, Stanford 1.12, 2009.

[19] A. Bermingham and A.Smeaton, Classifying sentiment in microblogs: is brevity an advantage?. Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM, pp. 1833–36, 2010. https://doi.org/10.1145/1871437.1871741

[20] S. Rude, E. M. Gortner, and J. Pennebaker, Language use of depressed and depression-vulnerable college students. Cognition and Emotion, pp. 1121–33. Taylor & Francis Group, 2004. https://doi.org/10.1080/02699930441000030.

[21] M. Park, C. Cha, and M. Cha, Depressive moods of users portrayed in Twitter. In Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD'12), pp. 1–8, 2012.

[22] M. De Choudhury, S. Counts, and E. Horvitz, Social media as a measurement tool of depression in populations. Proceedings of the 5th Annual ACM Web Science Conference, ACM. pp. 47–56, 2013. https://doi.org/10.1145/2464464.2464480.

[23] S. Malmasi, M. Zampieri, and M. Dras, Predicting post severity in mental health forums. In Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, pp. 133-137, 2016.

[24] B. Ay, O. Yildirim, M. Talo, U. B. Baloglu, G. Aydin, S. D. Puthankattil, and U. R. Acharya, Automated Depression Detection Using Deep Representation and Sequence Learning with EEG Signals. Journal of Medical Systems, 43(7), 1–12, 2019. https://doi.org/10.1007/s10916-019-1345-y

[25] C. Zucco, B. Calabrese, and M. Cannataro, Sentiment analysis and affective computing for depression monitoring. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, pp. 1988–95 (2017). https://doi.org/10.1109/BIBM.2017.8217966.

[26] N. Ramirez-Esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker, The Psychology of Word Use in Depression Forums in English and in Spanish: Texting Two Text Analytic Approaches. In ICWSM, 2008.

[27] P. G. F. Cheng, R. M. Ramos, J. Á. Bitsch, S. M. Jonas, T. Ix, P. L. Q. See, and K. Wehrle, Psychologist in a pocket: Lexicon development and content validation of a mobile-based app for depression screening. JMIR mHealth and uHealth, pp. e88–e88. JMIR Publications, 2016. https://doi.org/10.2196/mhealth.5284

[28] S. Almatarneh and P. Gamallo, A lexicon based method to search for extreme opinions. PloS One, pp. e0197816–e0197816. PUBLIC Library Science, 2018. https://doi.org/10.1371/journal.pone.0197816.

[29] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, Lexicon-based methods for sentiment analysis. Computational Linguistics - Association for Computational Linguistics, 37(2), 267–307, 2011. https://doi.org/10.1162/COLI_a_00049.

[30] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. Trends and Applications in Knowledge Discovery and Data Mining, 7867, 201–213, 2013. https://doi.org/10.1007/978-3-642-40319-4_18

[31] Z. Dong, and Q. Dong, HowNet - a hybrid language and knowledge resource. In: Proceedings of International Conference on Natural Language Processing and Knowledge Engineering, pp. 820–824, 2003.

[32] M. De Choudhury, S. Counts, and E. Horvitz, Predicting postpartum changes in emotion and behavior via social media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3267–3276, 2013.

[33] D. E. Losada and F. Crestani, A Test Collection for Research on Depression and Language Use. In International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 28–39. Cham: Springer, 2016.

[34] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, Effects of age and gender on blogging. In AAAI 2006 spring symposium on computational approaches to analysing weblogs, pp. 1-7, 2006.

[35] T. Davidson, D. Warmsley, M. Macy, and I. Weber, Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI Conference on Web and Social Media, 2017.

[36] M. Stankevich, I. Smirnov, N. Kiselnikova, and A. Ushakova, Depression Detection from Social Media Profiles. In Data Analytics and Management in Data Intensive Domains, pp. 181–194. Springer International Publishing, 2020. https://doi.org/10.1007/978-3-030-51913-1_12.

[37] L. Ma, Z. Wang, and Y. Zhang, Extracting Depression Symptoms from Social Networks and Web Blogs via Text Mining. Bioinformatics Research and Applications, 10330, 325–330, 2017. https://doi.org/10.1007/978-3-319-59575-7_29.

[38] S. K. Bharti and K. S. Babu, Automatic keyword extraction for text summarization: A survey. arXiv preprint arXiv:1704.03242, 2017.

[39] A. Dunne, M. Etropolski, A. Vermeulen, and P.Nandy, On Average: Data Exploration Based on Means Can Be Misleading. The AAPS Journal, pp. 60–67, us: Springer, 2012.

[40] C. Goutte, and E. Gaussier, A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. In European conference on information retrieval, pp. 345–359. Heidelberg, Berlin: Springer, 2005.

[41] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), pp.267-307, 2011.

[42] M. Taboada, Computational analysis of text sentiment, 2021. http://www.sfu.ca/~mtaboada/nserc-project.html.

[43] L. Zhang, R. Ghosh, M. Dekhil, M. Hsu, and B. Liu, Combining lexicon-based and learning-based methods for Twitter sentiment analysis. HP Laboratories, Technical Report HPL-2011, 89, 2011.

[44] P. Palanisamy, V. Yadav, and H. Elchuri, Serendio: Simple and Practical lexicon based approach to Sentiment Analysis. In Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013) (pp. 543-548), 2013, June.

[45] H. Christina, On Classification to and from Various Orders of Magnitude, 2008. https://serendipstudio.org/exchange/christina-harview/classification-and-various-orders-magnitude.