



A Comparative Study of Emotion Detection Methods Using NLP Techniques on Sentiment140 and IMDb Datasets

Bahar Asgari¹, Hamid Rastegari^{1,2,*}, Vahid Nejati³

¹ Department of Computer Engineering, Na.C, Islamic Azad University, Najafabad, Iran

² Big Data Research Center, Na.C, Islamic Azad University, Najafabad, Iran

³ Faculty of Psychology and Educational Sciences, Shahid Beheshti University, Tehran, Iran

Abstract: Emotional analysis plays a crucial role in identifying unknown aspects of human behavior, which nowadays is used by Artificial Intelligence techniques such as Natural Language Processing for the early prediction and diagnosis of mental illness. In recent years, computer science researchers have conducted numerous studies on detecting mental illness through computational methods. However, this domain still requires further exploration of unresolved challenges alongside existing solutions. This study reviews recent developments, key concepts, and issues in this area, focusing on two datasets: Sentiment140 and IMDb. Following preprocessing and normalization, several widely used classification algorithms such as Support Vector Machine, Logistic Regression, Random Forest, Extremely Randomized Trees, and Multi-Layer Perceptron are implemented, and their performance is compared based on accuracy. The experimental results indicate that the existence of a dictionary and the use of combined methods and algorithms such as Support Vector Machine and Logistic Regression on the IMDb dataset will achieve higher accuracy.

Keywords: Emotional Analysis, Mental Illness Detection, Logistic Regression, Support Vector Machine, Natural Language Processing .

1. Introduction

Emotion Analysis (EA) is a study field that tries to express the different people's emotions, behaviors and attitudes comparing their entities and features. Entity can be product, service, person or event. Emotion analysis is one of the most active research domains in natural language processing and it has also been studied widely in web mining and text mining. EA significance has increased at the same time of social media growth such as revisions, discussion forums, small weblogs, twitter and social networks. Now, for the first time in human history, a great amount of data have been registered in digital form for analysis [1] which are very practical for psychological sciences and detection of this field's diseases including depression i.e. the century disease. Nowadays, the extensive amount of information in networks let to recognize the emotion of text as an important area of study. Since detecting human emotions accurately is difficult, sometimes people use words that do not directly reflect their true emotions. Another point is that writing styles are not the same for everyone. Recognizing formal and informal writing styles is difficult. These are examples of factors that make emotional analysis less straightforward than it may seem at first glance [2]. Emotion refers to temperament external manifestation [3]. In general, the emotions are divided into two categories: basic and complicated [4]. The last common approach of computer-human interaction totally ignores the emotions while the emotional calculation results can have many benefits. Therefore, the first step is Human Emotion Recognition (HER) which was recently



paid more attention to. In HER, the collected information in this regard is similar to the human's symptoms of understanding others' emotions. Therefore, the ego is multi-dimensional in nature and includes textual and audio-visual features. It seems that text is the most common way under study because text processing is easier comparing other methods. However, user emotion analysis is a big challenge for human and machine. On the one hand, sometimes the people may not be able to recognize or express their emotions. On the other hand, the machines must have the appropriate accuracy for emotion modeling and also, they need advanced ML algorithms to develop emotional models. The importance increase of emotion analysis and the growth of social media such as surveys, dialog forums and social networks happened simultaneously. After emotion analysis, it is the time for emotion mining which can be a branch of DM. Emotion mining tries to recognize the existence of human emotions like happiness or fear from text, voice and visual artifacts of humans. The humans can show a great amount of emotions like sadness, happiness, anger, fear, etc. in their opinions [°].

Emotion mining is a field of emotional computing; more precisely, this field follows emotion mining based on people's observations and actions that can be obtained using their writings, speech, intonation, gestures, etc. Emotion analysis of each of these media is a special field of study. In this study, we focus only on text-based emotional equations. Textual emotion mining is a general concept including more accurate tasks that are Emotion Detection and Emotion Classification[°].

In fact, the response to a detection problem is "yes or no" i.e. if there is emotion expression in the text or not. However, the aim is responding to a classification problem of emotions' precise kinds (happiness, sadness, etc.) expressed in a text. In the emotional analysis, there are various types of datasets, some of the main types of which include: text, labeled, multi-purpose, multilingual, and temporal datasets. This diversity of datasets helps researchers train and evaluate emotional analysis models according to their specific needs. There are different methods and resources for collecting the required datasets for analysis, and below we introduce three categories of them;

- Dataset Based on Examining Responses
- Datasets Based on Self-declared Psychological Health Status
- Datasets Based on Post Interpretation
- Datasets Based on Membership in Associations

The existence of various datasets in emotional analysis and text mining helps researchers and experts in obtaining more accurate, extensive, and reliable results and gaining a better understanding of texts and their associated emotions. Additionally, the presence of different algorithms such as Support Vector Machines, Logistic Regression, Neural Networks, etc., which utilize machine learning techniques and natural language processing, significantly impacts emotional analysis. These algorithms can identify patterns, topics, and important aspects as well as extract emotions like anger, fear, happiness, etc. from texts through text content analysis. However, this task appears to be quite challenging.

This article is organized as follows: first, in Section 2, which covers related works. Then, in Section 3, experimental tests are discussed. Following that, Section 4 delves into the discussion of challenges, and finally in part 5, the conclusion has been presented.

2. Related Works

Emotion analysis which is considered as a branch of text mining is now the theme of the day and the research in this regard has been started since 2000. In the following, we examine the related works in each of categories.

2.1 Rule-based Methods

The designed rules in text classification system show the current "lexical", "semantic" and "morphological" patterns in textual data. Furthermore, these rules can show the current "semantic" or "phonological" aspects in textual data. Strapparava et al. used lexical method in 2004 and created a lexicon called WordNet Affect to show emotional and carnal information selecting and labeling a sub-set of Wordnet with passive and carnal concepts such as emotion, trait, feeling, etc.



[٧]. Another lexicon called WordNet Domains was also used to make it. WordNet Domains is a multi-lingual version of WordNet that provides at least one domain label (such as politics, sport, etc.) [٨].

The Valence Aware Dictionary and sentiment Reasoner (VADER) is a lexicon and rule-based method that determines sentiment scores from web-based media [9]. It assigns a sentiment intensity to each word in a sentence and then aggregates all intensities to obtain a score for the overall sentiment. VADER can discern whether a word is positive, neutral, or negative based on its lexicon of sentiment-laden words.

Many studies have been done based on rule-based method and since this method has a specific and pre-defined pattern and lexicon, it can be superior in some aspects comparing other methods; however, lack of comprehensive lexicon in every language in emotion analysis and different rule processing in different languages are considered as disadvantages of this method.

2.2 Machine Learning

Text mining systems based on machine learning are capable of learning from the educational samples and must change the educational data into templates which can be read by the machine. The input of ML algorithms for text classification is textual data which are vectorized and labeled by appropriate classes and they are the output of textual data classification models. One of the well-known textual data vectorization methods is a technique called Bag-of-Words.

Most of the conducted research has been conducted on English language. The first study on Persian language was done by Mohammadreza Shams et al. in 2012 [١٠]. Recently, emotion detection algorithms have also been introduced. Masoud Garshasbi et al. in a research published in Journal of Information Technology Management addressed emotion detection algorithms in Persian language. They pointed out that emotion detection of a sentence is not very accurate just considering an emotional word because there are other factors such as aggravators and deniers which affect the text's emotion. Therefore, they proposed an algorithm which was written in six ways considering different features. The results of implementing these algorithms on two data samples show that regarding more features, the algorithm accuracy will increase in which the part of speech factor has the highest effect [١١]. Mohammad Ehsan Basiri et al. addressed examining emotion extraction from Persian data of SPerSent in their study. They proposed the word set of CNRC which accompanies two labels. Furthermore, they used Nave Bayes algorithm for assessing their results and the results of the simulation showed that this algorithm has a better function for polarity detection than rank prediction [١٢]. In another study conducted in 2019, they addressed identifying depressed users in online forums and they have reached better function and higher F1 using combination of time and network features [١٣].

Generally, we can say that in the field of ML-based analysis and classification, two sets of experiment and test data are needed. We face the phenomenon of continuous improvement because the accuracy and efficiency of Machine Learning algorithms increase while they experience more and that gives them the possibility of better decision making; thus, they act well in examining multi-dimensional, multi-variable, structured and unstructured and they can do that in dynamic or unknown environments. However, in Persian language, using this technique has some problems such as lack of sufficient resources, lack of algorithm proper function on Persian unstructured texts, etc.

2.3 Hybrid

"Hybrid" systems combine the rule-based systems with ML-based systems. These systems are considered as each other's complement and they increase the accuracy of produced results. Many studies in the field of emotion analysis used this method. For example, Sergio G. Burdisso et al. [١٤] addressed in a study the early detection of depressed people. They used MapReduce technique and introduced a model called SS3 for classification which is costly in terms of computing but has a very good capability of explaining and overcomes the ERD problems.



Moreover, another study conducted in 2019 addressed depression detection using social networks and minor emotions production [15]. Lieva and colleagues [16] conducted a project aimed at assisting in the detection of early signs of depression through social media posts. The authors used emotional analysis as a feature extraction method using the VADER emotional analysis tool. VADER is completely equivalent to Textblob as both categorize sentiment polarity as positive, neutral, or negative in the same way. A sentence can be determined as having positive or negative sentiment based on the highest polarity score. From this project, the authors conducted an experiment to compare the scores when machine learning models are combined with VADER and without VADER. The results indicate that there is no significant difference when using VADER plus and not using it. This depends on the machine learning models.

We can say that hybrid method could solve the challenges of using each method alone as it has used two rule-based and ML methods simultaneously. A summary of some conducted studies has been shown in Table 1.

Table 1: A summary of some conducted studies							
Ref	Dataset		Metric	Outcome	Achievement	Data Mining	Using Method
	Platform	N					
[11]	Lms. iauec. ac. ir	100	Acc.	Identify Emotions	84% Acc.	Unsupervised	Nave Bayes
[12]	Dijikala	15000	Acc. P, R, F1	CNRC & Spresent	Solving the problem of lack of resources at the sentence level in Persian	Hybrid	Nave Bayes
[13]	Reddit	160	F P R ERDE ₀	Depress Non-depress	Early detection of depression. computationally expensive.in SS3 High F.	Unsupervised	SS3
[14]	Twitter	140 character limit	similarity	Positive words & negative words For depression	Sentiment of words & stress scores & scored efficiently	Hybrid	LDA & TensiStrength & SentStrengt h
[15]	CEASE	2539	Mean Recall (MR)	Depression, SEntiment, Emotion	MR of 56.47% on the emotion recognition task, 1.98% greater than the single-task equivalent model	Hybrid	Deep Learning
[16]	Questionnaire	511	F1 Acc.	Depression Detection	Highest score is achieved by AdaBoost Classifier with a 0.66-F1 score	Hybrid	extra Tree, AdaBoost, RF NB, LR



3. Experimental tests

In this research investigation, emotional analysis has been conducted using Natural Language Processing on two datasets using text classification methods and machine learning models, and emotional analysis techniques have been tested. Their accuracy has been calculated using formula (1). Fig (1) illustrates the overall framework of this research. In this experiment, two datasets, Sentiment 140 with 1.6 million tweet [21], and IMDB [22] with 50000 movie reviews ,have been utilized.

$$Accuracy = \frac{true\ negatives + true\ positives}{total\ examples} \quad (1)$$

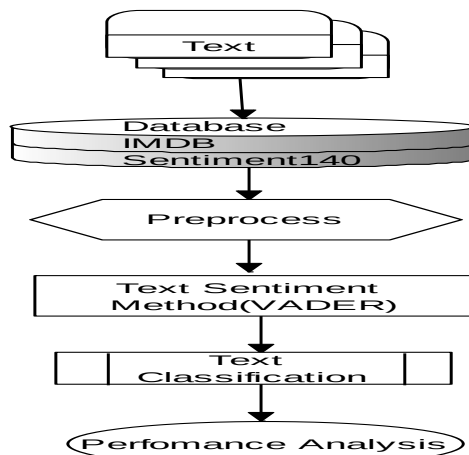


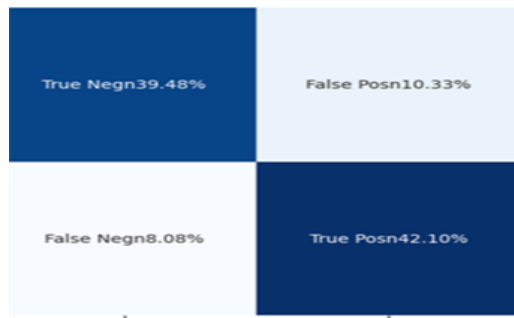
Fig 1- framework of the method used

3.1. Preprocessing

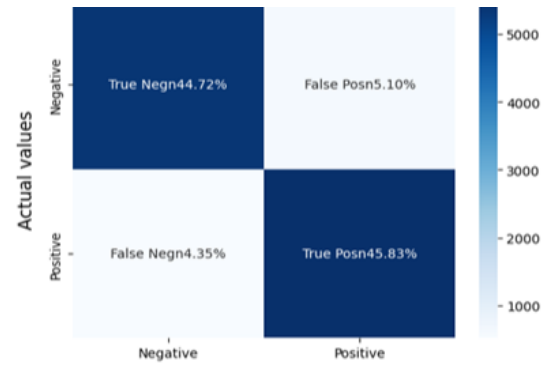
The datasets need to undergo preprocessing steps using Natural Language Processing tools before feature extraction and classification. First, unnecessary symbols, numbers or symbols, repetitive characters, URLs, and stop words are removed. Then, the texts pass through the tokenization process, dividing the dataset into small tokens. Lemmatization and stemming operations are performed. After that, the preprocessed data are split into 70% training sets and 30% test sets. For emotional analysis, we use the Valence Aware Dictionary for Sentiment Reasoning library (VADER), which is a rule-based tool.

3.2 Results and Analysis

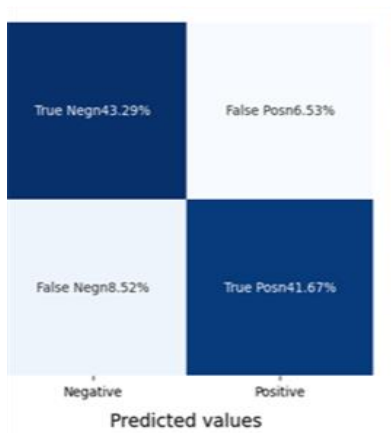
First, we use Term Frequency Inverse Document Frequency (TF-IDF) to extract features and then implement our classification models and evaluations. Additionally, we use K-fold cross-validation with K set to 2. We calculate the accuracy of a model based on confusion matrices. Fig 2 and 3 display confusion matrices for each of the classification methods for the IMDB and Sentiment 140 datasets.



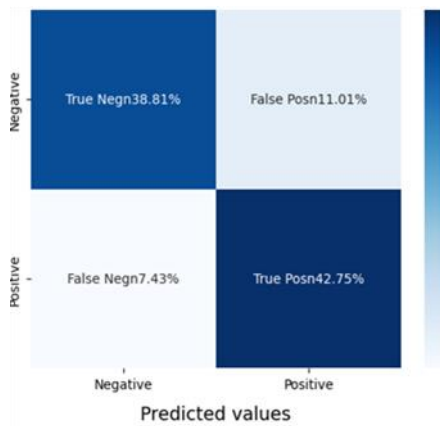
(a).Extra Tree



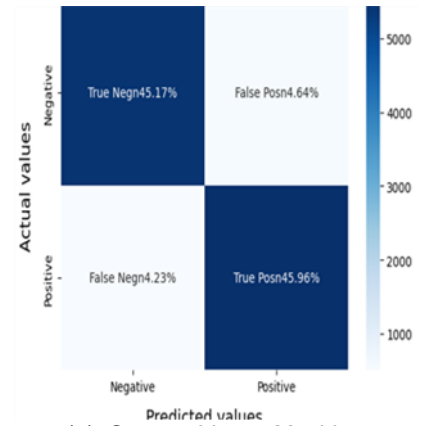
(b).Logistic Regression



(c). Multilayer Perceptron

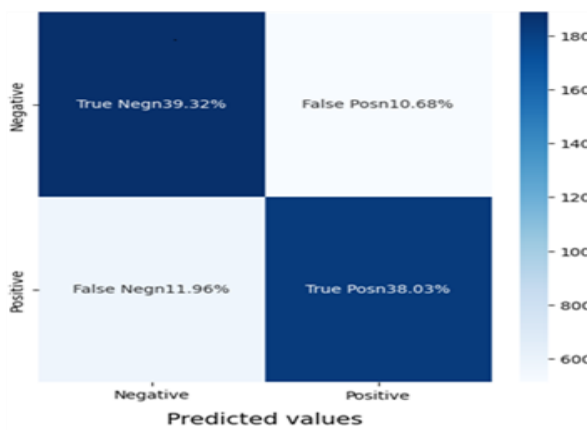


(d). Random Forest

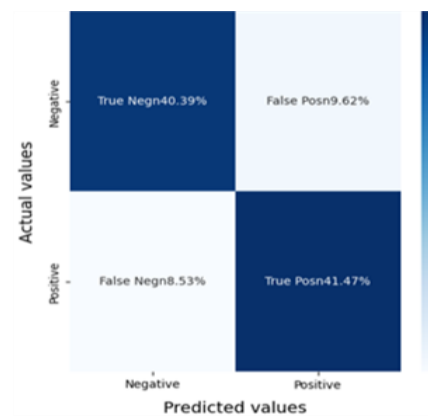


(e). Support Vector Machine

Fig 2- Confusion matrices of algorithms in IMDB dataset



(a). Extra-Tree



(b). Logistic Regression

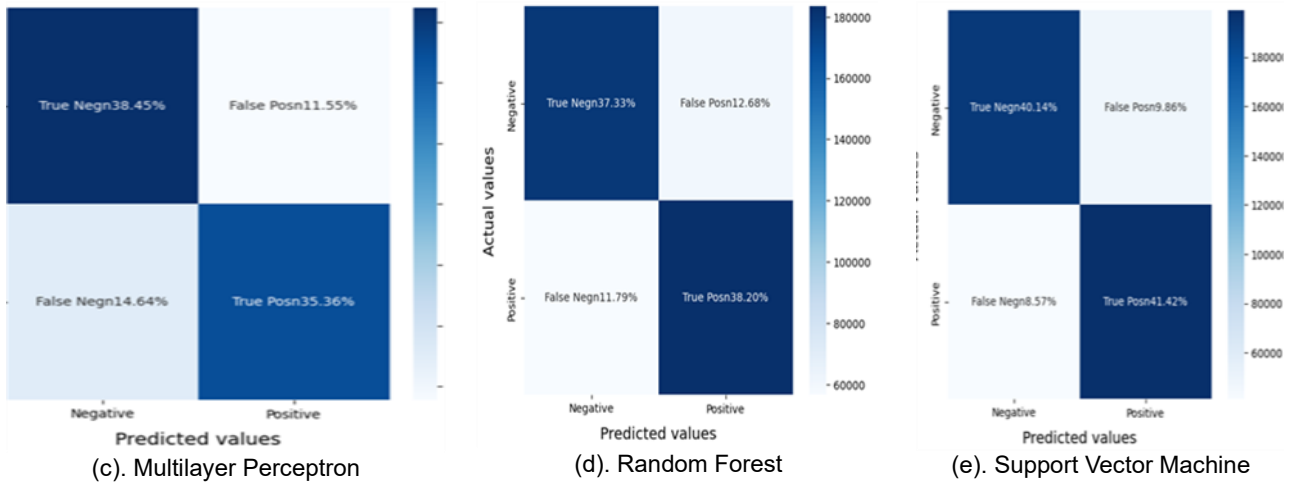


Fig 3- Confusion matrices of algorithms in Sentiment140 dataset

3.3 Criticism and general comparison of dataset collection methods

Most of the various studies in the past have experimented with natural language processing of texts from social media posts or Google Forms questionnaires. Each of the existing methods has its advantages and disadvantages. There are many methods and screening tools available in the field of psychology, and although they are very reliable and valid, they are also very costly on a large scale. In this study, we attempted to test two datasets with the same algorithms. The results shown in the table 2 indicate that in the IMDB dataset, all proposed algorithms had higher accuracy compared to Sentiment140, even though the number of samples in Sentiment140 is higher. Additionally, the SVC and LR algorithms also had higher accuracy than the others in both datasets.

Table 2- Comparison of Accuracy in two datasets		
Algorithm	IMDB	Sentimen140
SVC	0.91	0.82
LR	0.91	0.82
Random Forest	0.82	0.76
Extra Tree	0.81	0.77
MLP	0.81	0.77

4. Discussion

In emotion analysis, having a detailed and complete dataset is crucial for achieving high accuracy in analysis. Additionally, the methods of collecting these data and the features of each dataset aid in emotional analysis. In this research, experiments have been conducted on the Sentiment140 and IMDB datasets. The IMDB dataset is used for emotional analysis and reviewing movie reviews. This dataset contains user reviews about movies and is typically divided into two categories: positive and negative reviews. The IMDb dataset includes the following features: 1. Review, 2. Sentiment or Label. However, the Sentiment140 dataset contains user tweets. Each record in this dataset includes the features, that is, polarity, ID, data, user, and text.

In this research, the accuracy of all implemented algorithms in the IMDB dataset performs better, despite the presence of more features in Sentiment140. However, in emotional analysis on text, discrete features such as ID, Date, User, etc., do not substantially affect the classification accuracy. Additionally, in the IMDB dataset, because the text are extracted from movie reviews, there is more coherence in the text regarding a specific topic, which contributes to a higher classification accuracy. One of the challenges is that an opinion can be expressed by all the sentence components. for instance:



Noun: I enjoy watching these shoes.

Adjective: the shoes are beautiful.

Verb: I love these shoes.

Adverb: these shoes are really of good quality.

However, the text writers discuss different topic aspects in comparative sentences. The studies about text classification generally focus on adjectives and adverbs to detect their orientation. The most commonly used tool for adjective detection is WordNet. Claire Nedellec (1999) proposed an ML system that can obtain the relationship between classification and verb group based on input syntactic rules [۲۳]. Applying NLP tools in the psychiatric research field is also challenging now. The current NLP tools are sensitive in two aspects: First, the capability of NLP to reduce the complexity of texts without structure; second, the capability of understanding the relationships among words or concepts in an appropriate way. Because of this limitation, the NLP systems are now only able to provide rudimentary analysis, and it is only in psychiatry that the patients are often described by delicate concepts in terms of emotions or personality. Furthermore, the NLP tools have been designed almost exclusively for text exploration in English. For other languages, there are no tools, or if they exist, they can be used for very basic analysis. The limitations of NLP methods should be identified; however, the increasing progress of NLP can result in solving the challenges.

5. Conclusion

This article discusses emotional analysis concepts, introduces some datasets, compares two datasets, examines the existing challenges, and presents some related work.

In the present study, two common datasets were examined and experimented with using several classification algorithms. We observed that while the Sentiment140 dataset, which includes texts extracted from tweets along with other features and has a larger number of data points, exhibits fewer coherence and consistency compared to the IMDB dataset, which contains texts extracted from movie reviews. This difference in coherence and consistency partly affects the classification accuracy. Furthermore, it can be concluded that the presence of a vocabulary and a detailed dataset, along with the use of combined methods and efficient algorithms such as Support Vector Machine and Regression, as well as ensemble algorithms, increases the accuracy of Emotional analysis. Additionally, employing K-fold Cross Validation has had a positive impact on the results and has increased accuracy in large datasets like Sentiment140 compared to when this variable was not present. Emotional analysis can be also be used in marketing and trade; however, using it in psychology is an interesting topic and provides a great amount of research possibilities. Many other diseases can be detected using text mining and emotion analysis in which natural language processing and ML developments have helped a lot. It is proposed that in future, the psychological health of patients suffering from hemophilia or MS is examined using current approaches with social network data.

6. References

- [1] B. Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, no. 1, pp. 1-167, 2012.
- [2] S. S. Sadeghi, H. Khotanlou, and M. Rasekh Mahand, "Automatic persian text emotion detection using cognitive linguistic and deep learning," *Journal of AI and Data Mining*, vol. 9, no. 2, pp. 169-179, 2021.
- [3] H. I. Kaplan, and B. J. Sadock, *Synopsis of psychiatry: Behavioral sciences clinical psychiatry*: Williams & Wilkins Co, 1988.
- [4] S. Dhawan, K. Singh, D. J. I. J. o. I. Sehrawat, and C. Technology, "Emotion mining techniques in social networking sites," vol. 4, no. 12, pp. 1145-1153, 2014.
- [5] B. Liu, "Sentiment Analysis and Subjectivity," 2010.
- [6] A. G. Shahraki, "Emotion Mining from Text," University of Alberta, 2015.
- [7] R. Valitutti, "WordNet-Affect: an Affective Extension of WordNet," In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1083--1086, 2004.
- [8] B. Magnini, and G. Cavaglia, "Integrating Subject Field Codes into WordNet." pp. 1413-1418.
- [9] C. Hutto, and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text."



- [10] M. Shams, A. Shakery, and H. Faili, "A non-parametric LDA-based induction method for sentiment analysis." pp. 216-221.
- [11] M. Garshasbi, A. Rais-Rohani, and M. Kabaranzadeh Ghadim, "Presenting a Text Mining Algorithm to Identify Emotion in Persian Corpus," *Journal of Information Technology Management*, vol. 10, no. 2, pp. 375-389, 2018.
- [12] M. E. Basiri, and A. Kabiri, "Sentence-level sentiment analysis in Persian." pp. 84-89.
- [13] A. Shrestha, and F. Spezzano, "Detecting depressed users in online forums." pp. 945-951.
- [14] S. G. Burdisso, M. Errecalde, and M. J. E. S. w. A. Montes-y-Gómez, "A text classification framework for simple and effective early depression detection over social media streams," vol. 133, pp. 182-197, 2019.
- [15] M. E. Aragón, A. P. López-Monroy, L. C. González-Gurrola, and M. Montes, "Detecting depression in social media using fine-grained emotions." pp. 1481-1486.
- [16] V. Leiva, and A. Freire, "Towards suicide prevention: early detection of depression on social media." pp. 428-436.
- [17] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "A text classification framework for simple and effective early depression detection over social media streams," *Expert Systems with Applications*, vol. 133, pp. 182-197, 2019.
- [18] X. Tao, R. Dharmalingam, J. Zhang, X. Zhou, L. Li, and R. Gururajan, "Twitter analysis for depression on social networks based on sentiment and stress." pp. 1-4.
- [19] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes," *Cognitive Computation*, pp. 1-20, 2022.
- [20] I. N. Joharee, N. N. W. N. Hashim, and N. S. M. Shah, "Sentiment Analysis and Text Classification for Depression Detection," *Journal of Integrated and Advanced Engineering (JIAE)*, vol. 3, no. 1, pp. 65-78, 2023.
- [21] R. B. Alec Go, Lei Huang, "Sentiment140," G. e. al., ed., 2009.
- [22] Col Needham, "IMDB," IMDb.com, ed., 1990.
- [23] j. noee, m. jalali, and r. ghaemi, "Opinion mining: A review of the work done," 2013.