






Depression Detection Using Deep Learning and Natural Language Processing Techniques: A Comparative Study

Francisco Mesquita¹ , José Maurício¹ , and Gonçalo Marques² 

¹ Polytechnic of Coimbra, ISEC, Rua Pedro Nunes, 3030-199 Coimbra, Portugal
{a2018056868, a2018056151}@isec.pt

² Polytechnic of Coimbra, ESTGOH, Rua General Santos Costa, 3400-124 Oliveira dos Hospital, Portugal
goncalo.marques@estgoh.ipc.pt

Abstract. Depression is a frequently underestimated illness that significantly impacts a substantial number of individuals worldwide, making it a significant mental disorder. The world today lives fully connected, where more than half of the world's population uses social networks in their daily lives. If we interpret and understand the feelings associated with a social media post, we can detect potential depression cases before they reach a major state associated with consequences for the patient. This paper proposes the use of natural language processing (NLP) techniques to classify the sentiment associated with a post made on the Twitter social network. This sentiment can be non-depressive, neutral, or depressive. The authors collected and validated the data, and performed pre-processing and feature generation using TF-IDF and Word2Vec techniques. Various DL and ML models were evaluated on these features. The Extra Trees classifier combined with the TF-IDF technique emerged as the most successful combination for classifying potential depression sentiment in tweets, achieving an accuracy of 84.83%.

Keywords: Depression Detection · Sentiment Analysis · Natural Language Processing (NLP) · Twint · Antidepressant

1 Introduction

Depression is an increasingly common illness around the world and one that is likely to increase soon. Some of the symptoms of depression may be restlessness, irritability, impulsivity, anxiety, palpitations, sadness, loss of energy, a sense of hopelessness and many more [26]. According to World Health Organization (WHO), 3.8% of the population suffers from this disorder which means that approximately 280 million people around the world live with depression [1]. In Europe, 6.38% of the population suffers from depression, ranging from 2.58% in

the Czech Republic and 10.33% in Iceland [16]. It is often difficult to diagnose mental illness, however, the growth and globalization of social network usage can help to reduce the number of cases that go unnoticed. Social networks play a key role and have a direct correlation with depression as suggested by Yoon et al. [31]. Over the past few years, there has been an increase in the number of people interested in studying and using machine learning (ML) algorithms to create medical decision support systems [24]. This is due to the great evolution that has taken place in the industry in terms of computing power and the ever-increasing amount of data available [25].

The sentiment behind social media posts should be examined in order to diagnose depression as quickly and accurately as possible [6]. To achieve this, it needs a system capable of processing, analyzing, and deriving knowledge from a diverse and unstructured data set. One specific domain within the field of ML, particularly deep learning (DL), has the ability to accomplish this very task. Natural language processing (NLP) is capable of understanding human language and taking valuable information from it [32]. This has been a hot topic in research in recent years using data mining and ML techniques. The potential that these techniques could have for clinical use is very high as shown in the Ricard et al. study [23]. Consequently, it is proposed an ML approach to detect the sentiment associated with a tweet made on the Twitter social network. This could be the depressive, neutral, or non-depressive sentiment. Twitter was chosen because of its massive use, being the third most popular social network in the world. It also has a simple data model and an easily accessible API to collect data.

This paper objective is to create a predictive model capable of detecting a possible depressive feeling associated with a tweet. This will allow worldwide improvements in the way potential people at risk of depression are detected. The main contributions are: (i) Describe a full ML pipeline that covers collecting data, processing it, training a classification algorithm, and evaluating its performance, (ii) Comparative analysis of different feature generation techniques and classification algorithms, (iii) Compare the achieved results and proposed model with prior research works in the literature.

This paper has the following structure: Sect. 2 presents the summary of all similar papers in the literature, Sect. 3 is the methodology used, including information about the dataset, pre-processing techniques, label validation, exploratory data analysis, how we generate features with TF-IDF and Word2Vec, experimental setup and how we evaluate the models created. Section 4 shows all the results that we obtain with both ML and DL models. Section 5 presents the discussion where we state the findings of this study and make a comparison with what has been done previously in the literature. Section 6 are the conclusions we can draw from our work including contributions, limitations, and future work.

2 Related Work

In the last few years, several works have been proposed on how to automate the diagnosis of depression in a patient, to reduce the number of cases of depres-

sion that are increasing and reducing the number of suicides caused by major depression.

The study proposed by [7] used the public dataset Sentiment 140 which contains data without the presence of signals of depression. Adding to that data, they gathered a dataset with signals of depression, through the collecting data of Twitter with recourse to the Twint tool. The following keywords were used to pick tweets with the signals of depression: hopeless, lonely, antidepressant and depression. In pre-processing, the stop words, punctuation marks and hyperlinks were removed, and authors used Lemmatization to group different forms of the same word. After preprocessing the data, a Feature Generation was performed based on techniques such as Tokenization to separate the words in the text into a form that the machine understands. Valence Aware Dictionary and sEntimentReasoner (VADER) were also used to extract the polarity of the tweets to get the overall emotion of the text and finally Word2Vec was utilized to transform the text into word vectors.

After these data preparation processes, the dataset was divided into 60% for training and the rest was divided into validation and testing. They proceeded to classify the data using two types of approaches: a) a Long Short-Term Memory (LSTM) network; b) a hybrid CNN-LSTM model. In the first approach, they were able to obtain 90.33%, 91%, 91%, and 91% of Accuracy, F1-Score, Precision and Recall, respectively. On the other hand, in the second approach, they were able to get 91.35%, 91%, 92% and 91% on the same metrics, respectively.

Another study proposed by [9] achieved an improvement in these results using a combination of Word2Vec and DL models. It achieved 99.02% Accuracy, 99.04% Precision, 99.01% Recall and 99.02% F1-Score for the LSTM network. And, obtained 99.01% of Accuracy, 99.20% of Precision, 99.01% of Recall and 99.10% of F1-Score for the hybrid CNN+LSTM model.

Many works make use of feature extraction tools such as Bag-Of-Words (BOW), Tokenizer and TF-IDF models. In the study proposed by [29], the authors aimed to predict whether the person was not depressed, was half depressed, moderately depressed or severely depressed, so they used the unsupervised K-Means clustering algorithm to label the tweets. Decision Trees, Random Forest and Naive Bayes algorithms were used in this study for the classification of the tweets. The dataset was split into 80% for training and 20% for testing. In the end, they evaluated the performance of the algorithms through the classification metrics Accuracy, F1-Score, Recall, Precision and R-Score. The combination of TF-IDF models to generate features with the Random Forest algorithm stood out from this approach, having obtained 95% of Accuracy, 95% of Precision, 95% of Recall, 67% of R-Score and 95% of F1-Score.

The authors of [27] used a public dataset with 43000 tweets. For each tweet, a pre-processing was performed which consisted of removing non-alphabetic characters (e.g., HTML tags, punctuation, hashtags, numeric values, special characters, URLs), normalization the tweeters converting the text to lowercase, removing stop words (e.g., prepositions, conjunctions, and articles), and at last applied stemming. Since ML algorithms cannot process the raw text, TF-IDF was used to extract features to then be provided as input to the model. As a result, using

Multinomial Naive Bayes they achieved 72.97%, 74.58% and 75.04% accuracy, precision and recall, respectively.

Alsagri et al. [5] used almost the same pre-processing steps but the data was obtained through the Twitter API and was a much smaller amount, about 3000 tweets. It is also a differentiated approach as it tries to classify the user himself as depressive through the various tweets associated with him. Using TF-IDF it obtained 82.50% accuracy, 73.91% precision, 85% recall, 79% f1 score and finally, 77.50% AUC.

Kabir et al. [14] proposed a new topology to diagnose depression disease in Twitter messages called DEPTWEET and introduced a unique dataset labelled, with clinical validation, and for each label, a confidence score was assigned. The Twitter messages were retrieved using the Twint tool and the search keywords were defined based on the PHQ-9 questionnaire for depression. They classified each tweet as one of the four possible values: non-depressed, mildly depressed, moderately depressed, or severely depressed. As classifiers, the authors used Support Vector Machine (SVM), Bidirectional LSTM (BiLSTM), and two pre-trained transformer-based models: BERT and DistilBERT. ROC score was chosen as the evaluation metric and the best result was obtained using the transformer-based models, with the DistilBERT standing out and getting 78.88% in non-depressed, 74.72% in mildly, 78.79% in moderate and 86.60% in severe depression.

When compared to the literature, our work proposes a collection of tweets using the TWINT tool and the assignment of a label (POSITIVE, NEGATIVE, and NEUTRE) to the text through the sentence polarity score achieved by VADER. In addition, we will perform a manual validation of each phrase present in the data to ensure that the classes are assigned correctly, thus reducing the probability of error in the label assignment process done in the [14, 28] work. Word2Vec and TF-IDF were also used for feature generation. Finally, to classify the text, we use two DL architectures (LSTM and hybrid CNN+LSTM) and several ML algorithm.

3 Methodology

The methodology proposed in this article to detect depression consists of 3 steps: i) collecting data from Twitter using the Twint tool, cleaning and categorizing the collected tweets; ii) manual validation of the label assigned to each tweet and data augmentation; iii) generating features for each sentence using TF-IDF and Word2Vec to train ML and DL algorithms. As shown in Fig. 1.

3.1 Dataset

The data collection for this work consisted in acquiring Twitter data with signs of depression, using the Twint tool. The keywords lonely, depressed, frustrated, hopeless and antidepressant were used to obtain the phrases with signs of depression [7].

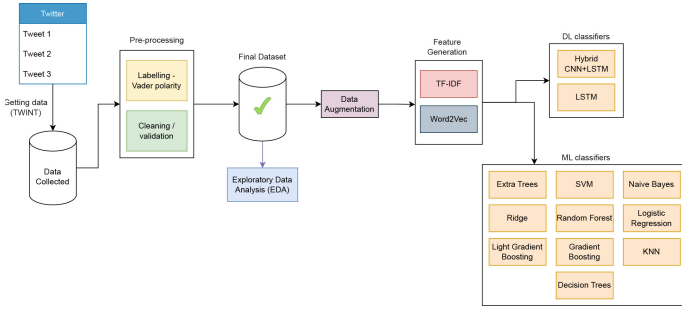


Fig. 1. Experimental setup.

Using Twint, we configure the search parameters such as the respective search keyword, the tweet limit, and the language in which the tweets are written. After that, the Twint tool will systematically extract the desired data from Twitter, aggregating tweets, user details, and even interaction metrics. At the end of this process, we have a dataset ready to be processed and analyzed.

All tweets were gathered on November 3, 2022. Twint tool starts by retrieving the latest tweets and continues to fetch older tweets until it reaches a stopping condition, such as a specified number of tweets or a certain time limit. In our case, it stopped when it reached the limit of 3500 collected tweets. Each word originated a dataset of tweets with 3500 records that were later converged to a final dataset with 17500 records.

3.2 Data Pre-processing

When collecting tweets for analysis, it is crucial to account for the presence of noise resulting from the limitations of the collection process, which is based on a single keyword. Where there is no control over the content of the tweets obtained. Therefore, pre-processing techniques were applied to the tweets. First, the sentences were normalized to convert them to lowercase. Next, all hyperlinks, hashtags, identifications of other users and emojis were removed. In addition, all stopwords were removed from the tweets collected through the stopwords function of the nltk library. These steps have been suggested in several previous works [11, 15, 22].

After the collected data had been cleaned, the VADER tool was used to categorize the tweets into Positive, Negative and Neutral. VADER is a tool widely used in sentiment analysis tasks due to its simplicity, effectiveness, and ability to handle domain-specific and colloquial language. It was introduced by Hutto et al. in 2014 [13] and has been employed in various applications, including social media monitoring, customer feedback analysis, and opinion mining [3, 8, 10, 12, 20]. To do the labelling of the tweets the compound score value generated by VADER was used as a base. It consisted in assigning the label Positive for compound score values greater than or equal to 0.05, the label Negative for compound score values less than or equal to -0.05 and values between -0.05 and 0.05 would

be identified as Neutral. These compound values are those recommended in the article by Hutto et al. [13].

During pre-processing, a check was made for the existence of missing values, and they were removed as recommended by [21]. With the removal of the missing values, 1057 records were lost out of the 17500 records in the final dataset. Besides the missing values, it was verified the existence of duplicate data. As suggested by [30] this resulted in the removal of 955 records.

3.3 Manual Validation of Label

The manual validation process for the sentences in the collected dataset involved excluding sentences with fewer than three words after pre-processing. This was done because we cannot determine the sentiment from a sentence of less than 3 words [18]. This step resulted in 1520 sentences being eliminated.

After this validation, the resulting sentences were checked to ensure that the label assigned by the VADER algorithm was correct or not. In instances where there was ambiguity in interpreting the pre-processed sentence, we turn to the corresponding original sentence to better understand its meaning and decide. When incorrect labels were identified, appropriate corrections were made to ensure accuracy.

Furthermore, as part of this procedure, a check was performed to ensure that the collected sentences were relevant to the theme of the study. In cases where the semantics of the sentences were incorrect or they were in a different language, they were discarded. As a result, a total of 10,920 sentences were subjected to validation. Out of these, 8,519 sentences were removed from the dataset, leaving 2,512 sentences that were used for the study. This high number of discarded sentences is due to several factors: (i) many of the collected tweets did not fit the topic of depression, often consisting of reviews, opinions, quotes or other types of the text unrelated to a depressive feeling; (ii) despite the Twint settings, some tweets came in other languages and were therefore removed; (iii) a few tweets were ambiguous or even contradictory about the possible associated sentiment and we decided to discard them.

3.4 Exploratory Data Analysis

During the exploratory data analysis, the purpose was to examine the distribution of the assigned labels before and after manual validation and dataset augmentation. Figure 2 illustrates the class distribution prior to validation and augmentation, indicating an imbalance among the classes. From the graph analysis, it is evident that a class imbalance exists, as VADER tends to assign a strong Negative sentiment label to sentences containing negative keywords. However, Fig. 3 presents the class distribution after the dataset went through validation and augmentation, revealing that the data is now almost perfectly balanced.

Furthermore, we built three Word Clouds that allow us to visualize which words are present in the collected sentences. The Word Cloud with the words of the sentences that were considered positive and negative are shown in Fig. 4 and

Fig. 5, respectively. Finally, Fig. 6 shows the words of the sentences that were considered neutral.

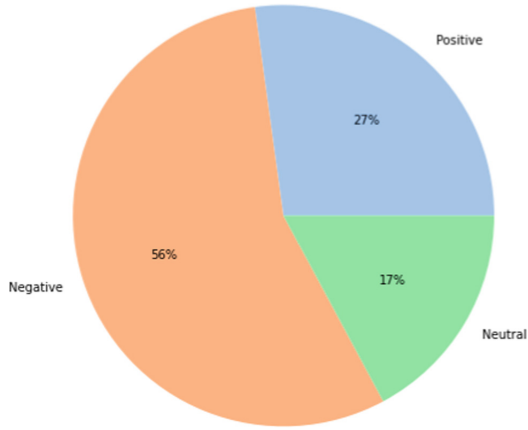


Fig. 2. Distribution of the classes before the manual validation and augmentation.

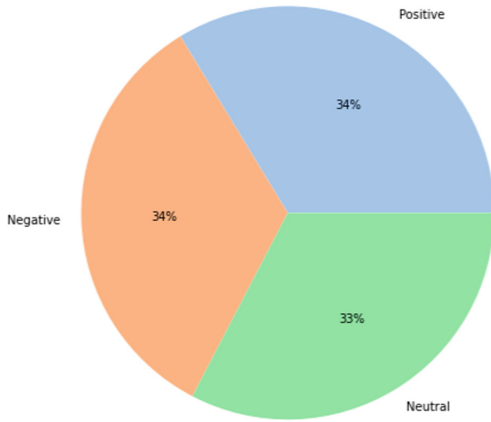


Fig. 3. Distribution of the classes after the manual validation and augmentation.

3.5 Feature Generation

The algorithms selected to generate features applying the algorithms described in the literature are presented in this section. Therefore, this work used the TF-IDF algorithm which consists of the vectorization of documents to calculate a score for each word based on its importance in the document and corpus [2].

Word2Vec algorithm allows representing all the words of the sentences extracted from Twitter in embeddings based on the similarity of the words [4]. This algorithm was configured with 300 for the embedding size and 10 for the window size.

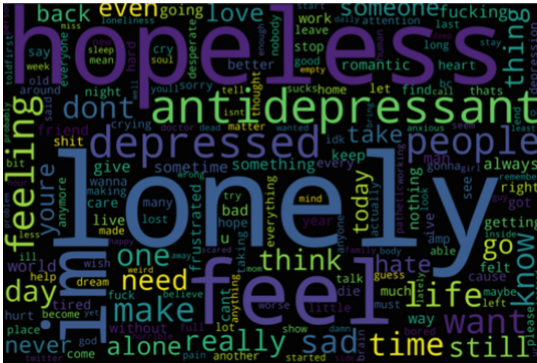


Fig. 4. Word Cloud with label Positive.



Fig. 5. Word Cloud with label Negative.

3.6 Experimental Setup

For the development of this work, the Pycaret library in version 2.3.10, TensorFlow in version 2.11.0, Scikit-learn in version 1.2.0, NLKT in version 3.7 and Gensim in version 3.6.0 were used. In addition to these libraries, the nlpaug library was also used in version 1.1.11 to do data augmentation. To increase the number of instances and diversity when training the ML algorithms, data augmentation was used. The augmented sentences were generated by changing certain words by synonyms throughout the sentence, always maintaining their meaning and the associated sentiment. In addition, this step allowed the target to be balanced. From 2512 sentences we now have in the final dataset 5032 instances.

4 Results

Only the top outcomes from the experiments conducted will be provided in this section. Therefore, Table 1 has presented the top 10 best ML algorithms with the features generation being performed by the Word2Vec algorithm. Table 2 presents the top 10 best ML algorithms, where the TF-IDF algorithm was used to generate the features.

Table [3–4] are presented the best results obtained with the DL algorithms using TF-IDF and Word2Vec algorithms to generate features, respectively. In bold are marked the best result.

Table 1. Top 10 ML algorithms with Word2Vec.

Models	Accuracy	Precision	Recall	F1-Score	AUC
Extra Trees Classifier	0.5775	0.5821	0.5775	0.5779	0.7430
Extreme Gradient Boosting	0.5444	0.5448	0.5444	0.5439	0.7383
Random Forest	0.5417	0.5454	0.5417	0.5419	0.7279
Light Gradient Boosting	0.5411	0.5434	0.5411	0.5403	0.7281
Quadratic Discriminant Analysis	0.5391	0.5434	0.5391	0.5383	0.7172
Gradient Boosting	0.5060	0.5112	0.5060	0.5045	0.6980
K-Neighbors Classifier	0.4921	0.4953	0.4921	0.4920	0.6826
Linear Discriminant Analysis	0.4735	0.4748	0.4735	0.4730	0.6540
Ada Boost	0.4728	0.4738	0.4728	0.4687	0.6448
Naive Bayes	0.4477	0.4597	0.4477	0.4335	0.6298

Table 2. Top 10 ML algorithms with TF-IDF.

Models	Accuracy	Precision	Recall	F1-Score	AUC
Extra Trees Classifier	0.8483	0.8501	0.8483	0.8487	0.9515
SVM	0.8086	0.8232	0.8232	0.8231	0.0000
Ridge Classifier	0.8086	0.8087	0.8086	0.8086	0.0000
Random Forest	0.8060	0.8089	0.8060	0.8066	0.9292
Logistic Regression	0.7556	0.7562	0.7556	0.7557	0.8935
Naive Bayes	0.7417	0.7583	0.7417	0.7415	0.8088
Decision Tree	0.6960	0.6966	0.6966	0.6961	0.7719
Extreme Gradient Boosting	0.6927	0.6941	0.6927	0.6928	0.8564
Light Gradient Boosting Machine	0.6748	0.6756	0.6748	0.6750	0.8325
Gradient Boosting	0.6305	0.6339	0.6305	0.62975	0.8081

By analysing the results, is concluded that the Extra Trees Classifier algorithm combined with TF-IDF with 84.83%, 85.01%, 84.83% and 84.87% of Accuracy, Precision, Recall and F1-Score, respectively. It is the most accurate solution to predict whether a person has depression, through the sentences collected from

Twitter. However, it is observed that the SVM algorithm describes a near performance with 80.86%, 82.32%, 82.32% and 82.31% in the same metrics.

Based on these results, we can conclude that the DL algorithms do not demonstrate a good performance to predict depression in people through the gathered tweets. It has the combination of the hybrid model with Word2Vec demonstrating the higher result with: 52.05% of Accuracy, 59.25% of Precision, 52.15% of Recall, 51.92% of F1-Score and 1.0313 Loss.

Table 3. Results of the DL algorithms with TF-IDF.

Models	Accuracy	Loss	Precision	Recall	F1-Score
LSTM	0.3305	1.0986	0.1102	0.3333	0.1656
Hybrid model	0.3285	1.0986	0.1095	0.3333	0.1648

Table 4. Results of the DL algorithms with Word2Vec.

Models	Accuracy	Loss	Precision	Recall	F1-Score
Hybrid model	0.5205	1.0313	0.5925	0.5215	0.5192
LSTM	0.4642	1.0882	0.4968	0.4595	0.4031

5 Discussion

This comparative study between ML and DL algorithms uses different feature-generation techniques. By examining how people express themselves on social media, we can, with a certain degree of confidence, determine if they are feeling depressed or not. It serves as a benchmark for future research in sentiment analysis, offering a methodology that collects raw tweets and processes those sentences to predict the sentiment that person intends to express. In addition, it demonstrates the use of data augmentation tools associated with NLP problems.

Based on the findings of this study, it is clear that using the Extra Trees Classifier along with the TF-IDF feature generation technique achieves good prediction results. On the other hand, it is verified through the same results that the DL algorithms were inferior to the ML algorithms, in both combinations. Except that when using the Word2Vec algorithm, the hybrid model was able to obtain a better performance than algorithms such as Naive Bayes, Ada Boost, Gradient Boosting, K-Neighbors Classifier and Linear Discriminant Analysis. This can be explained by the low amount of data available and its high complexity [19]. Although this has been verified, we believe that DL techniques have the potential for better performance with a larger and more diverse dataset.

The best result presented in this study proves to be superior to the results obtained by a pre-trained model based on a transformer [14]. Where the authors collected data from Twitter, through the Twint tool and the validation of the sentences was done by a doctor, as well as the keywords selected were based on questionnaires previously performed. On the other hand, in studies that used deep learning algorithms the results obtained were significantly higher than those demonstrated in this work [7,9]. However, it was not possible to know the structures of the algorithms and their configurations to justify the results obtained.

In addition, most literature works used a larger dataset than the one used in this work. Either they used a public dataset such as sentiment 140, which is already labelled and in some cases medically validated, having a larger number of instances [27]. Or in other studies, data is collected following the same methodology as this study but is also used a validated public dataset to increase the number of tweets and to balance target [7,9]. Whereas, the authors of this study collected sentences from Twitter that were validated regarding the sentiment expressed in it by themselves.

The use of a two-step validation process, involving VADER initially and subsequent manually, enhances the accuracy of the sentiment associated with the phrase, thereby increasing confidence in the obtained results, despite the potential bias associated with the authors' interpretation. Also, it is possible to verify that the Extra Trees Classifier algorithm presents an AUC of 95.15%, which means that our ML algorithm has a good ability to distinguish between Positive, Negative and Neutral classes.

Table 5 presents a comparison between the proposed method and previous works in the literature.

When evaluating the insights provided by the literature, it remains unclear whether the algorithms' impressive performance translates into effective class distinction. We can take our study as an example, although the SVM and Ridge classifier algorithms exhibit strong predictive capabilities for sentiment analysis on tweets, a closer examination of the AUC value reveals their inability to effectively differentiate between the classes within the dataset.

This work shows that it is possible to classify sentences from Twitter according to the associated sentiment, doing so in sentences where there are many grammatical gaps, the vocabulary is not homogeneous and often there are both spelling and grammatical errors. According to the authors, utilizing raw data comprising colloquially written phrases that closely reflect real-life experiences enhances the classifier's performance and adds greater significance to the results. This approach is seen as more valuable compared to models trained on transformed phrases without spelling or grammatical errors, and homogeneous, failing to capture the authentic social media reality.

Table 5. Comparison of the results with the literature.

Ref	Model	Data	Labelling	Pre-processing	Performance
[7]	Hybrid CNN-LSTM	- Sentiment 140 non-depressive public data; - 1.6 million tweets; - Gathered depressive data with Twint which has no quantity information	- All non-depressive phrases came from public data; - Depressive sentences are the one gathered; with Twint; - No validation	- Remove hyperlinks, digits and stop words; - Text case change and Slang substitution; - Spell checking and lemmatization; - Feature generation with Word2Vec	Accuracy: 91.35%, Precision: 92%, Recall: 91% and F1-Score: 91%
[9]	Hybrid CNN-LSTM	- Random non-depressive tweets obtained from Kaggle; - Depressive tweets derived from Twint; - No information about quantity	- Random tweets from public data treated as not depressive; - Depressive data gathered with Twint; - No validation	- Remove links, images and URLs; - Remove punctuation and stop words; - Stemming and lemmatization; - Tokenization	Accuracy: 99.01%, Precision: 99.20%, Recall: 99.01% and F1-Score: 99.10%
[14]	DistilBERT	- Collected data using Twint; - final data contains 41191 tweets	- Human annotators; - Manual validation from expert psychologist	- WordPiece tokenizer	AUC: 79.75%
[29]	Random Forest	- Data collected from Twitter API, Kaggle and using Twint; - 16000 tweets where 8000 are negative and 8000 are positive	- Automatic annotation using K-means clustering	- Remove links and punctuation; - Feature extraction using Bag of Words TF-IDF and Tokenizer	Accuracy: 95%, Precision: 95%, Recall: 95% and F1-Score: 95%
[27]	Multinomial Naive Bayes	- Public data collected from Kaggle - 43000 tweets	- Already present on data; - No validation	- Emoji extraction and slung substitution; - Remove links, timestamp, digits, symbols, proper nouns and stop words; - Spelling correction and lemmatization; - Feature extraction with Bag of Words	Accuracy: 72.97%, Precision: 74.58% and Recall: 75.04%
[5]	Linear SVM	- Collected data manually and using Twitter API; - About 3000 tweets	- Manual human validation of the depressive tweets; - Non-depressive tweets were collected randomly and without validation	- Tokenization and Stemming; - Normalization: turn to lower case and remove links, emojis, symbols, mentions, retweets and punctuation; - Feature extraction using TF-IDF	Accuracy: 82.50%, Precision: 73.91%, Recall: 85%, F1-Score: 79% and AUC: 77.50%
-	Proposed method: Extra Trees	- Collected data manually using Twint - Data augmentation - 5032 tweets	- Two Step labelling; - Auto labelling with VADER; - Manual validation of each sentence sentiment	- Normalization: turn to lower case and remove links, hashtags, mentions, emojis and stopword; - Remove less than 3 words sentences; - Remove sentences that were not in English; - Feature extraction with both TF-IDF and Word2Vec	Accuracy: 84.83%, Precision: 85.01%, Recall: 84.83%, F1-Score: 84.87% and AUC: 95.15%

6 Conclusion

Depression is a highly common illness in our society, characterized by feelings of sadness, lack of interest, and potential psychological and physical harm. Individuals with depression tend to engage more with social media compared to those without the condition. Detecting the underlying emotions expressed in social media posts could aid in identifying and monitoring individuals who require mental health support, ultimately enhancing their well-being.

Throughout this work, a predictive model capable of predicting whether a given Twitter phrase has a negative, neutral, or positive sentiment was developed. The best DL model was the Hybrid combination (CNN + LSTM) with Word2Vec achieving a low accuracy value (52.05%). Overall, the best model created was the ML classifier Extra Trees combined with TF-IDF achieving 84.83% of accuracy. Therefore, it concludes that Extra Trees Classifier with TF-IDF is the best combination to predict a possible depressive feeling associated with a sentence.

Nevertheless, this work has several limitations. The dataset size is limited which can make it very difficult to create a model with the ability to generalize to external examples. Furthermore, future research will be needed to see if the model can maintain this performance on external data. Also, our validation of sentences will always have a bias associated with what may be the interpreta-

tion of reviewers, and which may not correspond to the real feeling behind the sentence. The use of Large Language Models (LLMs) can automate and improve the data labelling process leading to potentially better classifier performance.

On Future work, we will try to figure out how to generalize the model created to external data. The augmentation technique used can lead to a potential bias where despite considerably increasing the volume of data, we still have a low variance, which can affect the model's performance. To solve this, multiple different augmentation techniques and classifiers can also be analyzed to improve the results.

References

1. Depression. <https://www.who.int/news-room/fact-sheets/detail/depression>. Accessed 26 Oct 2022
2. TF-IDF for Document Ranking from scratch in python on real world dataset. <https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089>. Accessed 09 Jan 2023
3. Al-Garaady, J., Mahyoob, M.: Public sentiment analysis in social media on the SARS-CoV-2 vaccination using VADER lexicon polarity (2022)
4. Almeida, F., Xexéo, G.: Word embeddings: a survey (2019). <https://doi.org/10.48550/arXiv.1901.09069>
5. Alsagri, H.S., Ykhlef, M.: Machine learning-based approach for depression detection in twitter using content and activity features. *IEICE Trans. Inf. Syst.* **E103.D**(8), 1825–1832 (2020). <https://doi.org/10.1587/transinf.2020EDP7023>
6. Babu, N.V., Kanaga, E.G.M.: Sentiment analysis in social media data for depression detection using artificial intelligence: a review. *SN Comput. Sci.* **3**(1), 74 (2021). <https://doi.org/10.1007/s42979-021-00958-1>
7. Bhargava, C., Al, E.: Depression detection using sentiment analysis of tweets. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **12**(11), 5411–5418 (2021)
8. Biswas, S., Ghosh, S.: Drug usage analysis by VADER sentiment analysis on leading countries. *Mapana J. Sci.* **21**(3) (2022)
9. Dessai, S., Usgaonkar, S.S.: Depression detection on social media using text mining. In: 2022 3rd International Conference for Emerging Technology (INCET), pp. 1–4 (2022). <https://doi.org/10.1109/INCET54531.2022.9824931>
10. Elbagir, S., Yang, J.: Sentiment analysis on twitter with Python's natural language toolkit and VADER sentiment analyzer. In: *IAENG Transactions on Engineering Sciences*, pp. 63–80. WORLD SCIENTIFIC (2019). <https://doi.org/10.1142/9789811215094.0005>
11. Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P.: Study of twitter sentiment analysis using machine learning algorithms on Python. *Int. J. Comput. Appl.* **165**, 29–34 (2017). <https://doi.org/10.5120/ijca2017914022>
12. Hossain, M.S., Rahman, M.F.: Customer sentiment analysis and prediction of insurance products' reviews using machine learning approaches. *FIIB Bus. Rev.* (2022). <https://doi.org/10.1177/23197145221115793>
13. Hutto, C., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, pp. 216–225 (2014). <https://doi.org/10.1609/icwsm.v8i1.14550>

14. Kabir, M., et al.: DEPTWEET: a typology for social media texts to detect depression severities. *Comput. Hum. Behav.* **139**, 107503 (2023). <https://doi.org/10.1016/j.chb.2022.107503>
15. Kolchyna, O., Souza, T.T.P., Treleaven, P., Aste, T.: Twitter sentiment analysis: lexicon method, machine learning method and their combination (2015). <https://doi.org/10.48550/arXiv.1507.00955>
16. Arias-de La Torre, J., et al.: Prevalence and variability of current depressive disorder in 27 European countries: a population-based study. *Lancet Publ. Health* **6**(10), e729–e738 (2021). [https://doi.org/10.1016/S2468-2667\(21\)00047-5](https://doi.org/10.1016/S2468-2667(21)00047-5)
17. Macrohon, J.J.E., Villavicencio, C.N., Inbaraj, X.A., Jeng, J.H.: A semi-supervised approach to sentiment analysis of tweets during the 2022 Philippine presidential election. *Information* **13**(10), 484 (2022). <https://doi.org/10.3390/info13100484>
18. Mendon, S., Dutta, P., Behl, A., Lessmann, S.: A hybrid approach of machine learning and lexicons to sentiment analysis: enhanced insights from twitter data of natural disasters. *Inf. Syst. Front.* **23**(5), 1145–1168 (2021). <https://doi.org/10.1007/s10796-021-10107-x>
19. Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., Muharemagic, E.: Deep learning applications and challenges in big data analytics. *J. Big Data* **2**(1), 1 (2015). <https://doi.org/10.1186/s40537-014-0007-7>
20. Newman, H., Joyner, D.: Sentiment analysis of student evaluations of teaching. In: Penstein Rosé, C., Martínez-Maldonado, R., Hoppe, H.U., Luckin, R., Mavrikis, M., Porayska-Pomsta, K., McLaren, B., du Boulay, B. (eds.) AIED 2018. LNCS (LNAI), vol. 10948, pp. 246–250. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-93846-2_45
21. Prakash, T.N., Aloysius, A.: Data preprocessing in sentiment analysis using twitter data. *Int. Educ. Appl. Res. J.* **3**, 89–92 (2019)
22. Ramadhani, A.M., Goo, H.S.: Twitter sentiment analysis using deep learning methods. In: 2017 7th International Annual Engineering Seminar (InAES), pp. 1–4 (2017). <https://doi.org/10.1109/INAES.2017.8068556>
23. Ricard, B.J., Marsch, L.A., Crosier, B., Hassanpour, S.: Exploring the utility of community-generated social media content for detecting depression: an analytical study on Instagram. *J. Med. Internet Res.* **20**(12), e11817 (2018). <https://doi.org/10.2196/11817>
24. Shailaja, K., Seetharamulu, B., Jabbar, M.A.: Machine learning in healthcare: a review. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 910–914. IEEE (2018). <https://doi.org/10.1109/ICECA.2018.8474918>
25. Sidey-Gibbons, J.A.M., Sidey-Gibbons, C.J.: Machine learning in medicine: a practical introduction. *BMC Med. Res. Methodol.* **19**(1), 64 (2019). <https://doi.org/10.1186/s12874-019-0681-4>
26. Tiller, J.W.G.: Depression and anxiety. *Med. J. Aust.* **199**(S6), S28–S31 (2013). <https://doi.org/10.5694/mja12.10628>
27. tweets, Hemanthkumar, Latha: Depression detection with sentiment analysis of tweets. *Turk. J. Comput. Math. Educ.* (2019)
28. Wani, M.A., ELAffendi, M.A., Shakil, K.A., Imran, A.S., El-Latif, A.A.A.: Depression screening in humans with AI and deep learning techniques. *IEEE Trans. Comput. Soc. Syst.* (2022). <https://doi.org/10.1109/TCSS.2022.3200213>
29. Woods, C., Adedeji, M.: Classification of depression through social media posts using machine learning techniques. *Univ. Ibadan J. Sci. Logics ICT Res.* **7**(1), 19–28 (2021)

30. Yadav, N., Kudale, O., Rao, A., Gupta, S., Shitole, A.: Twitter sentiment analysis using supervised machine learning. In: Hemanth, J., Bestak, R., Chen, J.I.Z. (eds.) *Intelligent Data Communication Technologies and Internet of Things. Lecture Notes on Data Engineering and Communications Technologies*, pp. 631–642. Springer, Cham (2021). https://doi.org/10.1007/978-981-15-9509-7_51
31. Yoon, S., Kleinman, M., Mertz, J., Brannick, M.: Is social network site usage related to depression? A meta-analysis of Facebook-depression relations. *J. Affect. Disord.* **248**, 65–72 (2019). <https://doi.org/10.1016/j.jad.2019.01.026>
32. Zhou, B., Yang, G., Shi, Z., Ma, S.: Natural language processing for smart healthcare. *IEEE Rev. Biomed. Eng.*, 1–17 (2022). <https://doi.org/10.1109/RBME.2022.3210270>