Comparative Analysis of Machine Learning and Deep Learning Algorithms for Twitter-Based Depression Detection

Emanoel Faria dos Santos
Dept. of Computer Science
Federal Institute of Southeast Minas Gerais
Rio Pomba, Brazil
emanoelvisali@gmail.com

Lucas Grassano Lattari
Dept. of Computer Science
Federal Institute of Southeast Minas Gerais
Rio Pomba, Brazil
lucas.lattari@ifsudestemg.edu.br

Maurício Archanjo Nunes Coelho
Dept. of Computer Science
Federal Institute of Southeast Minas Gerais
Rio Pomba, Brazil
mauricio.coelho@ifsudestemg.edu.br

Bianca Portes de Castro
Dept. of Computer Science
Federal Institute of Southeast Minas Gerais
Rio Pomba, Brazil
bianca.castro@ifsudestemg.edu.br

***Abstract*—Depression is a significant mental health disorder affecting millions of individuals globally. Detecting depressive symptoms from written texts, especially on social media platforms like Twitter, has received considerable attention. In this paper, we present a comparative analysis of machine learning and deep learning algorithms for depression detection on Twitter. We propose an innovative approach that integrates a multi-layer Long Short-Term Memory (LSTM) architecture with a Multi-Head Attention component. Our approach achieves up to 99% across all key metrics, including accuracy, recall, F1-score, and precision. However, it should be noted that these high scores are obtained in certain instances, thus being highly competitive compared to other relevant works. Despite facing challenges such as imbalanced datasets and user-annotated data, these remarkable results mark a promising advancement in the field of text-based depression detection.**

***Index Terms*—depression detection, twitter, social media, machine learning, deep learning, sentiment analysis, lstm, multi-head attention.**

I. Introduction

Depression, as defined by the World Health Organization [1], is a persistent mood disorder affecting approximately 280 million people globally. It often results in a loss of interest in activities and presents a considerable diagnostic challenge due to its diverse symptoms and associated social stigma. The COVID-19 pandemic, marked by extended periods of isolation, has exacerbated these issues, reinforcing the urgency to tackle depression effectively. This crisis underscores the crucial need for innovative, scalable, and accessible tools and strategies to detect and mitigate depressive symptoms, promoting mental health wellness beyond these difficult times.

Considering these concerns, there has been a significant increase in research focused on the possibility of detecting signs of depression in patients' written texts. Social media platforms, in particular, have been recognized as a prolific ground for such studies, given their wealth of user-generated content that reflects personal thoughts, emotions, and experiences.

The proliferation of social media platforms has resulted in a vast accumulation of textual data. Among these platforms, Twitter is particularly advantageous for identifying indicators of depression in user posts. It serves as a significant tool to better comprehend, detect, and deal with depressive symptoms.

Twitter was chosen as the platform for analysis due to its unique characteristics and capabilities. Its homogeneous post volume, typically around 240 characters, simplifies the analysis. The platform's real-time, public nature allows for instantaneous evaluations and systematic data collection via Twitter's API. Moreover, available metadata such as geographical location and posting time provide additional valuable context. Given its diverse user demographic, Twitter delivers a representative and generalizable dataset. Consequently, the wealth of diverse and nuanced data generated by users discussing their daily experiences and challenges presents a unique opportunity for the robust analysis of depressive symptoms.

The massive volume of available data requires automated approaches for efficient and accurate analysis and detection of these signs. Based on that, methods using natural language processing, sentiment analysis, machine learning, and deep learning have been shown to be capable of inferring individuals' mental states [2].

Gupta et al. [3] use machine learning algorithms to predict the progression of depression based on emotional, behavioral, and cognitive dimensions. The methodology involves the use of two publicly available datasets of positive and negative tweets. The study employs various machine learning models, including Decision Trees, Support Vector Machines (SVM), K-nearest Neighbor (KNN), and Long Short-Term Memory

(LSTM), alongside oversampling and undersampling techniques to address class imbalance in the dataset. The results demonstrate the superior performance of the LSTM model in depression detection, achieving the highest recall value of 0.75 and precision of 0.84. The study concludes that resolving class imbalance significantly enhances the performance of the psychological analysis model, with the SMOTE approach outperforming the RUS approach in depression detection, and the LSTM model delivering the highest accuracy of 0.82.

Nadeem et al. [4] propose a novel diagnostic approach to detect depressive sentiments through social media text. They curated a depression dataset, manually annotating it to encapsulate both implicit and explicit depressive and non-depressive tweets. The data underwent preprocessing and feature extraction using techniques such as TF-IDF, N-gram, and pre-trained word embeddings. The study employed various machine learning and deep learning algorithms on raw, binary, and ternary labeled data, culminating in the proposal of a unique deep-learning-based hybrid model with an attention mechanism. The results demonstrated the model's superior performance, significantly enhancing the accuracy of 97.4 for binary labeled data and achieving an accuracy and F1-score of 82.9 for ternary labeled data. The authors concluded that their method outperformed raw labels on real-time tweets, effectively capturing implicit depressive statements.

Amanat et al. [5] depict a strategy that utilizes a one-hot encoding methodology and the RNN-LSTM approach to characterize depressive symptoms from text data. The dataset used for this study was obtained from the Kaggle website, and preprocessing techniques such as stemming, lemmatization, and one-hot encoding with PCA were applied for data cleaning and feature extraction. The trained model, based on the RNN-LSTM approach, achieved an impressive accuracy of 99% with a reduced false positive rate. The evaluation results showcased the superior performance of the proposed framework, outperforming other methods such as Naive Bayes, SVM, CNN, and Decision Trees in terms of accuracy, precision, recall, and F1-measures.

In this paper, we propose a novel methodology for the detection of depressive signs in social media texts, particularly from the Twitter platform. We evaluate several typical machine learning algorithms, and a multi-layer Long Short-Term Memory (LSTM) architecture, enhanced with a Multi-Head Attention component, and benchmark its performance against three other studies. The principal findings of our work reveal that the LSTM model outperforms other evaluated techniques across multiple metrics, including accuracy, recall, F1-score, and precision.

This paper is organized as follows: Section II details our methodology, describing the data collection and preprocessing steps, the handling of data imbalance, the employed machine learning and deep learning algorithms, and our approach towards classification. Section III presents our experimental analysis, which includes a comparison of our method with three other studies, outlining our findings and insights. Section IV concludes the paper, highlighting our contributions and outlining directions for future research in this field.

II. Methodology

Our methodology was divided into the following stages: 1) acquisition of tweet datasets and user information (Section II-A); 2) pre-processing of data (Section II-B); 3) balancing of data classes (Section II-C); and 4) computation of classification methods (Section II-D). A comprehensive visual representation of these stages can be seen in Figure 1.

A. Data Collection

We used four distinct datasets, each available on Kaggle, a reputable open-source platform. These datasets were used as the basis for comparative analysis and for testing the adaptability of our method to different contexts.

The first, known as the "Twitter Depression Dataset" [6], comprises tweets extracted from the Twitter API. These tweets are categorized as either indicative of depression (d_tweets) or not (non_d_tweets), with a cleaned version also available. Despite the dataset's relevance, it possesses inherent biases due to the subjective selection of non-depressive tweets by its author. Moreover, there is a slight imbalance towards non-depressive tweets, with approximately 1,500 more in that category. Each entry provides a detailed set of tweet data, including ID, conversation ID, creation time, date, timezone, content, language, hashtags, among others.

On the other hand, the Sentiment140 Dataset [7] comprises 1.6 million tweets, all annotated for sentiment analysis. The polarity of sentiments is represented with labels (0 = negative, 2 = neutral, 4 = positive), determined by linking positive emoticons to positive sentiments, and analogously for negative sentiments. The dataset features columns for sentiment polarity (target), tweet id (ids), date, query (flag), user, and content (text).

The "Sentimental Analysis for Tweets" Dataset [8] is curated for sentiment analysis, particularly to detect signs of depression in social media language. It encompasses a singular file ("sentiment_tweets3.csv") providing tweet ID, content, and a binary label indicating depression signals presence (0 = no, 1 = yes). Although the dataset offers a diverse range of tweets supporting extensive analysis, the procedure for label assignment was not described. This dataset is licensed under GPL 2 and is not expected to receive future updates.

Lastly, the "DepressionTweets" Dataset [9] includes a training file comprising 30,068 tweets labeled as either depressive or non-depressive. However, the labeling method appears to be keyword-centric, focused on explicit statements of depression such as "depressed", "want to die", and "life is miserable". This approach can result in false positives (non-depressive tweets incorrectly marked as depressive) and false negatives (depressive tweets misidentified as non-depressive), thereby affecting the dataset's overall reliability. Despite these limitations, it offers a significant basis for examining the language associated with depression on social media. This dataset is not slated for future updates and is licensed under the Public Domain (CC0).

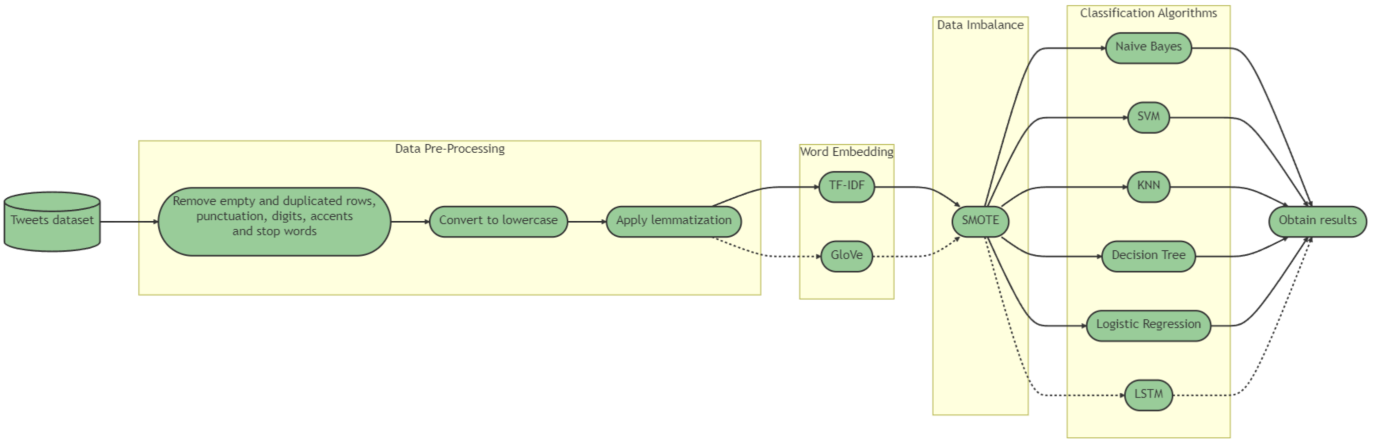

Fig. 1. This diagram illustrates our methodology, applied individually to each Twitter dataset, from initial pre-processing through to classification. Tweets undergo processing and lemmatization, then are transformed via TF-IDF and GloVe embedding (notably, the latter is used exclusively with LSTM, as indicated by the dashed line). SMOTE is utilized to address data imbalance before implementing several classifiers for depression prediction, including LSTM, which uniquely utilizes GloVe embedding. Lastly, results generated from all the employed classifiers are compared against other works.

B. Data Pre-Processing

The data pre-processing stage was conducted using Python 3.9, a popular programming language renowned for its simplicity and wide array of libraries facilitating data manipulation.

Several measures were taken, which included: eliminating empty and duplicate entries, removing punctuation and numerical digits, removing accents and various symbols, converting terms to lowercase, eliminating stop words, and applying lemmatization. These measures aim to standardize terms, minimizing noise and variability in the results.

The Pandas library was employed to handle data manipulation and analysis. This library provided DataFrames, a flexible data structure for handling structured data.

The NLTK (*Natural Language Toolkit*) library was another key component in our pre-processing stage. NLTK is a platform for building Python programs to work with human language data, offering tools for tasks such as classification, tokenization, stemming, tagging, parsing, semantic reasoning, and wrappers for NLP application. In our methodology, 'nltk' was used for stop word removal and lemmatization.

Additional libraries were used for specific tasks: 'string' library was employed to remove punctuation and numerical digits from the text data; 'unicodedata' library, part of Python's standard library, was used to remove accents and other diacritical marks from characters.

C. Data Imbalance and Classification Approach

Upon processing our data, we observed an imbalance between the number of entries with negative and positive polarity. To address this issue, we utilized the *Synthetic Minority Over-sampling Technique* (SMOTE), a proven method for balancing datasets [10]. SMOTE identifies samples from the minority class and generates new synthetic points similar to their nearest neighbors. This enhances the representation of the less-represented label and subsequently improves the classification performance. Upon the application of SMOTE, both classes held an equal number of samples.

We examined two key word-embedding techniques: *Term Frequency-Inverse Document Frequency* (TF-IDF) [11] and *Global Vectors for Word Representation* (GloVe) [12]. Prior to this, we trialed both Bag-of-Words and Word2Vec. However, given their comparatively suboptimal performance, we decided to concentrate our research efforts on the TF-IDF and GloVe techniques. The choice of these specific techniques was not only guided by their own merits but was influenced by their extensive utilization in other evaluated studies.

We employed TF-IDF encoding to transform words into numerical vectors. TF-IDF, although computationally less demanding, was juxtaposed with GloVe, which is known for its higher processing and memory requirements. In an attempt to balance computational efficiency and model performance, we aimed to evaluate whether the difference in outcomes between the GloVe and TF-IDF methods was substantial. This assessment was critical as it is beneficial to have the option of deploying a less computationally intensive model, particularly if the accuracy of such a model does not significantly differ from its more computationally demanding counterpart.

We also implemented *3-fold Cross-Validation* (CV) to assess the performance of the machine learning algorithms, considering varying splits of the training and testing set. The average of the CV metrics provides a robust and unbiased estimate of the model's performance on unseen data. The only exception to the application of CV was the Long Short-Term Memory (LSTM) model. For the LSTM architecture, overfitting issues were mitigated using a validation set and the Early Stopping technique instead of applying CV, due to the computational complexity of deep neural networks.

Before delving into the specifics of the classification algorithms, it's worth mentioning that, in addition to SMOTE, we also implemented the *Singular Value Decomposition* (SVD) on the continuous vectors generated by TF-IDF. This technique is used to reduce the dimensionality of sparse matrices, such as those typically produced by TF-IDF, whilst preserving the essential data structure. Although this may influence the evaluated metrics, it is anticipated to enhance computational performance.

D. Machine Learning and Deep Learning Algorithms

We applied the traditional machine learning algorithms in this methodology with standard hyperparameters, aiming to strike a balance between performance and computational efficiency. For the implementation of these algorithms, we adopted the scikit-learn (sklearn) library, a widely used machine learning library in Python. To preprocess the text data and create word embeddings, we implemented the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization technique.

We employed Multinomial Naive Bayes, a classifier that has seen extensive use in text processing tasks. The Support Vector Machine algorithm [14] implemented uses the radial basis function (RBF) kernel and allows a maximum of one thousand iterations. In the case of the K-Nearest Neighbors (KNN) classifier, we set the value of 'k' to three. The Decision Tree classifier [15] applies the Gini coefficient to determine node splits, and finally, the Logistic Regression model [13] sets a threshold of one thousand iterations as a stopping criterion.

We implemented a Long Short-Term Memory (LSTM) architecture [16], a specialized Recurrent Neural Network (RNN) variant designed to tackle the vanishing gradient problem. In traditional RNNs, long-range dependencies are difficult to learn due to vanishing or exploding gradients. LSTMs manage this issue through the use of gates: the forget gate, input gate, and output gate. These gates decide what information to discard, store, or output, respectively. Our implementation of LSTM encodes words into numerical vectors using the GloVe method, a pre-trained model with 840 billion tokens and 300-dimensional word vectors.

The model architecture features bidirectional layers, effectively capturing the dependencies in both forward and backward directions. A variable number of these bidirectional layers, along with variable units within each layer, allows the model to adapt to the complexity of different tasks. These architectural parameters, such as the number of layers, units, and other hyperparameters, vary depending on the specific work being compared. Additionally, the batch size utilized in the experiments is also adjustable.

To further enhance the model, we incorporated a multi-head attention component [17]. Originating from Transformer architectures, this mechanism enables the model to focus on different parts of the input while making predictions. Unlike single-head attention, multi-head attention enables the model to attend to multiple, possibly conflicting, contexts simultaneously. This makes it particularly effective at tasks where understanding context from different perspectives is crucial.

To address the problem of overfitting, we implemented a dropout rate of 20% on the units and used a validation set containing 5% of the training samples. We also introduced an *Early Stopping* mechanism, halting the training process if there is no reduction in validation loss after six epochs. Lastly, the training set contains 10% of the original samples.

III. Experimental Analysis

In this section, we make comparisons with three distinct studies, considering their respective datasets for depression detection, to ensure a fair comparison.

Similar to the methodology followed by Gupta et al. [3], we merged "sentiment_tweets3.csv" and "Sentiment140.csv" files into a single dataset. Only the tweet text and label columns were considered. The label (polarity of tweets) is either 0 (indicating negativity) or 1 (signifying positivity). Upon reading the data, a total of 1,610,314 records (tweets with their respective labels) were obtained. An imbalance between the number of words with negative and positive polarity was observed after the preprocessing step. After this step, we had 780,139 records labeled as 0 (depressive) and 784,991 records labeled as 1 (non-depressive). The Synthetic Minority Oversampling Technique (SMOTE), also used by Gupta et al. was employed to balance the dataset. After applying SMOTE, both classes had an equal number of 784,991 samples.

Comparing all algorithms listed in Section II, the Long Short-Term Memory (LSTM) model developed in this study showed superior recall and F1-Score metrics. This was anticipated due to the employment of GloVe for word encoding, bidirectional layers, and an attention component. These elements are considered state-of-the-art in the current literature, making them competitive, except against Transformer models. However, the LSTM model by Gupta et al., which uses TF-IDF and SMOTE, exhibited superior accuracy and precision. Detailed results can be viewed in Table I.

TABLE I
PERFORMANCE FOR THE ALGORITHMS USED IN THIS WORK AND GUPTA ET AL. [3], CONSIDERING DIFFERENT METRICS.

Classifier	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.76	0.69	0.73	0.74
SVM	0.64	0.50	0.53	0.60
KNN	0.61	0.62	0.62	0.61
Decision Tree	0.70	0.67	0.68	0.69
Logistic Regression	0.71	0.70	0.71	0.71
LSTM Imbalanced [3]	0.79	0.72	0.74	0.78
LSTM (SMOTE) [3]	**0.84**	0.75	0.79	**0.83**
LSTM (GloVe)	0.80	**0.80**	**0.80**	0.80
LSTM (GloVe and SMOTE)	0.80	**0.80**	**0.80**	0.80

It is important to note that the application of SMOTE to the LSTM model with GloVe and attention component did not cause significant difference in results. Given the disparity of 4,852 samples between the classes out of the total 1,610,314 records, the impact of SMOTE might be minimal. However, Gupta et al. report a significant difference when using TF-IDF with and without SMOTE. This discrepancy might be due to variations in the programming tools used in both studies or differences in hyperparameters, as the authors did not provide detailed information about their LSTM model. Our LSTM model was implemented with 3 bidirectional layers, 128 units for each layer, a dropout rate of 20%, and a learning rate of 0.003.

Despite this, both studies report a common finding: among the main known machine learning algorithms, LSTM-based approaches are more effective across all considered metrics.

We also conducted a comparative analysis with Amanat et al. [5], as depicted in Table II. In their research, Amanat et al. achieved near-perfect results by implementing a Long Short-Term Memory (LSTM) architecture with 2 layers and 60 units each. Moreover, they employed a one-hot encoding strategy and the Principal Component Analysis (PCA) algorithm for numerical encoding of the texts and dimension reduction. This approach is not very common as it tends to generate high-dimensional sparse vectors, whereas LSTMs typically handle sequences represented by dense and continuous vectors.

TABLE II
PERFORMANCE FOR THE ALGORITHMS USED IN THIS WORK AND AMANAT ET AL. [5], CONSIDERING DIFFERENT METRICS.

Classifier	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.73	0.56	0.63	0.68
SVM	0.60	0.65	0.62	0.62
KNN	0.70	0.61	0.65	0.68
Decision Tree	0.66	0.67	0.66	0.66
Logistic Regression	0.71	0.64	0.67	0.69
LSTM [5]	**0.98**	**0.99**	**0.98**	**0.99**
LSTM (GloVe and SMOTE)	0.87	0.86	0.86	0.86

The LSTM model we propose uses the same dataset as Amanat et al. for the results in Table II, and is similar to the LSTM described in the previous comparison. However, the architecture used consists of 2 layers with 64 units each, a dropout of 10% is applied to the units and a learning rate of 0.001. Like Amanat et al., we use 90% of the data for training and 10% for testing.

Another significant issue pertains to the imbalance of the original dataset, which contained 4,687 tweets labeled as 0 (non-depressive) and 3,082 labeled as 1 (depressive). While Amanat et al. mention having balanced the dataset, they do not provide details on how this was achieved. In our proposed methodology, we utilize SMOTE.

The difference in results is not very clear. Amanat et al. describe the use of 10-fold Cross-Validation with 100 epochs in each fold. Moreover, they do not mention the use of regularization techniques such as Early Stopping, L1, L2, or dropout. Similarly, they do not describe the use of a validation set. Hence, it can be hypothesized that overfitting may be present, despite the excellent results reported.

Lastly, we conducted a comparative analysis with the work of Nadeem et al. [4], as shown in Table III. The results found in our study and their work are very similar. Both utilize a LSTM architecture with 3 layers of 32 units, 10% dropout, and other components mentioned previously. Both works achieved near-perfect results, around 0.99 for all compared metrics. Even though SVM and Logistic Regression display identical values in the table, it's essential to highlight that this is due to rounding off. In our experiments, the LSTM model obtained the highest values in all four metrics, albeit by a slim margin.

TABLE III
PERFORMANCE FOR THE ALGORITHMS USED IN THIS WORK AND NADEEM ET AL. [4], CONSIDERING DIFFERENT METRICS.

Classifier	Precision	Recall	F1-Score	Accuracy
Naive Bayes	0.86	0.94	0.90	0.90
SVM	**0.99**	**0.99**	**0.99**	**0.99**
KNN	**0.99**	0.98	0.98	0.98
Decision Tree	0.98	0.97	0.97	0.97
Logistic Regression	**0.99**	**0.99**	**0.99**	**0.99**
LSTM (GloVe) [4]	-	-	0.97	0.97
LSTM (GloVe and SMOTE)	**0.99**	**0.99**	**0.99**	**0.99**

The method proposed by Nadeem et al. employs GLoVe-based encoding and a Deep Neural Network architecture that combines LSTM and CNN, Gated Recurrent Units (GRU), and an attention layer. Interestingly, the LSTM proposed in our study exhibits competitive results even when compared to a more complex architecture.

Our experiments further underscore that the use of LSTM architectures with multiple layers, in conjunction with GLoVe and multiple head attention components, yields stable and promising results across all assessed metrics, as compared to conventional methods.

In the future, we intend to make the source code of all mentioned experiments publicly available to facilitate their replication.

IV. DISCUSSION

A. Limitations

Although the results presented in Section III are robust and underscore the applicability of LSTM architectures, it's important to consider certain limitations associated with the problem of depression detection in social media texts.

One of the primary limitations is the lack of databases certified by mental health professionals. These databases are generally user-tagged or labeled based on specific inserted terms. We observed a shortage of quality databases with appropriate certification in this field, diminishing the impact of academic contributions aimed at solving this problem.

Another limitation pertains to the lack of temporal context in the data. Twitter posts are inherently dynamic and may exhibit temporal variations that our model does not account for. A tweet about depressive symptoms could possess different implications depending on its temporal context, an aspect not considered in our current methodology.

Furthermore, the absence of standardized methodologies and variability in results pose significant issues. Many published papers use their own databases, often without making them publicly available for comparison. Even when the dataset used in some studies can be located, they seldom provide essential details to replicate their methodologies, making a proper comparison challenging. There was also a noticeable absence of information explaining certain achieved results.

Another issue to consider is the adoption of deep artificial neural networks. Although they are considered state-of-the-art in recent periods, interpreting the decisions made by these networks is often impracticable. These networks are frequently referred to as "black boxes" in literature, posing challenges in healthcare areas where the explicit rationale behind the classification of texts as depressive or non-depressive is often unattainable.

B. Generalizability of Findings

To effectively interpret our study's outcomes, it is crucial to delineate its scope and limitations. This subsection focuses on evaluating the generalizability of our results, which are drawn from Twitter data.

Firstly, the cultural and linguistic context of the Twitter posts can significantly influence the applicability of our findings. The constraints imposed by Twitter, such as character limits and a tendency for informal language, may alter the linguistic features captured by our model. This raises questions about the model's performance when applied to texts from different languages or social platforms.

Secondly, Twitter is just one of several platforms where people share their thoughts and feelings. While the study sheds light on mood detection based on Twitter activity, it is unclear whether these conclusions can be extended to other platforms like Facebook, Instagram, or more specialized online forums.

Thirdly, the diversity of Twitter's user base is not entirely reflective of the general population. Some demographic groups might be underrepresented on Twitter, which calls into question the broader applicability of our findings.

Furthermore, the unstructured nature of Twitter data could present challenges for comprehensive mental health assessments. The absence of contextual information could be a limiting factor when translating these findings into a clinical setting.

Finally, our LSTM-based model excels in analyzing Twitter data, but its adaptability to other platforms or conditions might require changes. This could involve anything from data preprocessing adjustments to fine-tuning model parameters.

C. Challenges

In addition to the aforementioned limitations, researching the proposed problem presents further challenges. A major challenge is dealing with datasets that represent sensitive information of patients with mental disorders. Prioritizing the ethical handling of such data necessitates compliance with both individual consent and current data protection regulations. This necessitates compliance with legal frameworks such as the Brazilian Lei Geral de Proteção de Dados (LGPD) and the European General Data Protection Regulation (GDPR), which emphasize the privacy and autonomy of individuals over their personal data. It's crucial to anonymize the information to prevent the identification of individuals. Also, every aspect of data collection and storage should consider these legal and ethical considerations.

Additionally, the role of bias in such research needs examination. Depending on how data collection is performed, it might consider only specific groups from certain countries or regions, making generalizing findings to other contexts difficult. Therefore, publicly available databases must be properly audited by mental health professionals, considering the aforementioned ethical and privacy issues.

D. Concluding Remarks

This paper introduces a novel methodology for detecting depression signs in social media texts, with a specific focus on the Twitter platform. The method employed a multi-layer LSTM architecture, complemented with a Multi-Head Attention component. This approach, when benchmarked against three other studies, demonstrated competitive performance across multiple metrics such as accuracy, recall, F1-score, and precision.

Despite confronting various challenges, including the presence of unbalanced datasets and data annotated by users themselves, the LSTM-based method emerges as a promising solution, indicating substantial progress in the domain of depression detection in texts. The implementation of this approach facilitates the development of a computational system capable of continuously monitoring content posted on social networks, potentially identifying mood states of specific individuals and autonomously detecting profiles that may be experiencing depressive episodes.

Furthermore, it is of extreme importance that future research in this area focuses on the creation of databases that have been verified by mental health professionals. This would ensure the scientific validity of the findings and augment the relevance of the methods employed for the classification of depressive signs.

Lastly, it is important to explore and evaluate methodologies that employ Transformer-based algorithms, comparing their results with those of LSTM-based architectures, to continuously enhance and broaden the impact of this vital field of study.

References

[1] "Depressive disorder (depression)," World Health Organization, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression. [Accessed: 24-May-2023].

[2] A. L. Glaz et al., "Machine Learning and Natural Language Processing in Mental Health: Systematic Review," J. Med. Internet Res., vol. 23, no. 5, Art. no. e15708, May. 2021, doi: 10.2196/15708.

[3] S. Gupta et al., "Psychological Analysis for Depression Detection from Social Networking Sites," Comput. Intell. Neurosci., vol. 2022, Art. no. e4395358, Apr. 2022, doi: 10.1155/2022/4395358.

[4] A. Nadeem et al., "Depression Detection Based on Hybrid Deep Learning SSCL Framework Using Self-Attention Mechanism: An Application to Social Networking Data," Sensors, vol. 22, no. 24, Art. no. 9775, Jan. 2022, doi: 10.3390/s22249775.

[5] A. Amanat et al., "Deep Learning for Depression Detection from Textual Data," Electronics, vol. 11, no. 5, Art. no. 676, Jan. 2022, doi: 10.3390/electronics11050676.

[6] "Twitter Depression Dataset," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/hyunkic/twitter-depression-dataset. [Accessed: 17-April-2023].

[7] "Sentiment140 dataset with 1.6 million tweets," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/kazanova/sentiment140. [Accessed: 17-April-2023].

[8] "Sentimental Analysis for Tweets," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/gargmanas/sentimental-analysis-for-tweets. [Accessed: 17-April-2023].

[9] "DepressionTweets," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/samrats/depressiontweets. [Accessed: 17-April-2023].

[10] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[11] K. Sparck Jones, "A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL," J. Doc., vol. 28, no. 1, pp. 11–21, Jan. 1972, doi: 10.1108/eb026526.

[12] J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," in Proc. 2014 Conf. Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, Oct. 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[13] D. R. Cox, "The Regression Analysis of Binary Sequences," J. Roy. Stat. Soc. Ser. B (Methodol.), vol. 20, no. 2, pp. 215–242, 1958. [Online]. Available: https://www.jstor.org/stable/2983890. [Accessed: 28-May-2023].

[14] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proc. Fifth Annu. Workshop on Computational Learning Theory - COLT '92, New York, NY, USA, 1992, pp. 144–152, doi: 10.1145/130385.130401.

[15] J. R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[16] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[17] A. Vaswani et al., "Attention Is All You Need," arXiv.org, Jun. 12, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762v5. [Accessed: 28-April-2023].