Detecting Mental Health Issues Via Social Media

Abhishek Aryan
Dept. of Software Engineering
Delhi Technological University
Delhi, India
abhishekaryan_2k18se007@dtu.ac.in

Abstract—**While mental illness diagnosis rates have vastly improved over the last few decades, many cases remain undetected. This can be attributed to the fact that mental health is still a taboo topic that is not brought up widely, even though depression is a very common illness. Mental illness-related symptoms can be seen on Twitter, Facebook, Reddit, and other web forums, and as NLP techniques progress further, we can speed up and perform mass screening of these users to allow them to get help and provide other resources to them faster. The research points to these mental and verbal cues being prevalent and easily detectable in online speech on a platform like Twitter. Passive monitoring of these issues can significantly help in reducing the number of deaths and providing significant psychological help to needful users. This can even be used to supplement current screening procedures. Though this monitoring social media activity often leads to violation of user privacy, the inclusion of this along with other privacy features can help to combat this. More advanced methods can also be used to identify trolling and sarcasm in internet culture correctly.**

Keywords—**major depressive disorder, social media, analysis, mental health,**

I. Introduction

Depression is one of the most common mental ailments worldwide. It has affected more than 264 million people [1].

Depression is different from usual mood fluctuations, which people typically have, which are short-lived and temporary. Depression and more severe counterparts like Major Depression can result in severe impairments that hinder or restrain one's ability to carry out major life activities.

Even though it is so prevalent, depression is still a taboo topic in many parts of the world and thus does not receive adequate attention that is needed. More than 70% of people do not consult a doctor in the early stages of depression, which further deteriorates their condition.

Depression also leads to various other issues like anxiousness, feelings of low self-worth, suicidal tendencies, and many more. However, global provisions and facilities for the diagnosis, promotion, and treatment of mental disorders of this type have been deemed inadequate. Although 87% of the world's governments provide certain primary health care services to combat mental illness, 30% do not have initiatives, and 28% do not have a clearly defined budget for mental health [2]. In fact, for most types of mental illness, there is no accurate laboratory test to diagnose it properly; Typically, the diagnosis is based on self-identified perceptions of the patient, actions reported by family or associates, and mental status assessment.

As we increasingly integrate social media into our lives, we let more of our personas be available online. Twitter is a popular social media platform used by more than 330 million users worldwide and has around 130 million daily users. Compared to other social media like Instagram and Facebook, users can choose not to disclose any personal information so, users can express themselves without any tangible bounds (other than the character limit). As we will see, it is one factor that leads to more free speech on the platform and helps us see a franker and more accessible version of their users. They can easily tweet and post about their emotions, moods, and personal endeavors. This openness helps us peek under the veil at times and get an insight into their daily lives and mental headspace.

While this idea has been explored extensively, no dataset has been made for this cause, so we had to improvise and collect our own dataset. Several tools already exist for scraping, as analyzing tweets has become an increasingly prevalent method to learn and gather information about human behavior.

Once the user gets flagged by the system for heightened risk, a more thorough assessment can be carried out by a medical professional. The targeted individual could then be provided with more helpful resources helping them understand and manage their current condition. This also helps in monitoring at-risk users and help them. The methodology can also be augmented to be applied to images to detect severe self-harm, addiction, and various other problems. Around 23% of all deaths in the world are caused due to mental and substance abuse disorders. The inclusion of the risk factor can significantly help to reduce this percentage.

We use the mined data from Twitter in conjunction with a typical set of tweets via the "Sentiment140" dataset to compare the ground and control group. We start by using some simplistic Machine Learning techniques and compare them with a typical LSTM implementation. We would further like to see some attention and transformer models applied to this field. However, the usage of a Long Short Term Memory model (LSTM) already gives an excellent result.

II. Related Works

A. *Understanding the criteria:*

This project unifies different psychology and data science avenues. this means that we have rich literature and a plethora of related works in psychology, psychiatry, medicine, sociolinguistics, and data science.

Understanding our corpus and hence the psychology is an integral part of all the related works.

In Reference [3], Cloninger et al. explored how personality plays a significant undertaking in the onset and susceptibility to an episode of depression. They also explore how introverted people can be more affected by issues like low self-esteem and anxiety. A note concludes the paper on personality development and how it can lead to a reduction in future venerability to depression. It explores how the converse can also happen and how an episode of depression can change the user's traits and personality.

Other studies such as the one done by A.T. Beck et al. found that lack of social support and decreased self-esteem are significant factors associated with higher depression rates [4].

Another study conducted by A. M. Abdel-Khalek found that the association of somatic symptoms like tension, heart pain, insomnia, anorexia, migraine, and weight gains are also correlated to the onset of depression [5].

In sociolinguistics, studies like the one by T. E. Oxman et al. showed that linguistic speech analysis could identify and sort patients into groups based on paranoia and depression. Computerized analysis of this written corpus can also help in automated flagging and detection of such behaviors [6].

While studies have strengthened our understanding of factors associated with mental illness to date, a notable drawback of previous research is that it relies heavily on tiny, frequently homogeneous samples of people who may not actually be representative of the wider population. In addition, usually, these studies are focused on surveys, relying on retrospective mood self-reports and health observations: a process that restricts temporal granularity. That is, such evaluations are intended to collect high-level summaries over long periods of time about experiences.

It has been challenging to collect finer-grained longitudinal data, given the resources and invasiveness needed to track individuals' behavior over months and years.

B. *Analysis and Detection*

Moving on to more data-science related research, one of the first mental-health based assessments were done by S. Chellappan and R. Kotikalapudi via analyzing the web activity of college students [7]. Similarly, another study by M. A. Moreno et al. demonstrated how status updates on the social media platform Facebook could reveal symptoms of major depressive episodes [8]. Their control group was also based on college students.

TABLE I. GROUP DIVISION FOR A STUDY DONE IN COLLEGE STUDENTS

Division	**Categories**			
	Computer Science	***Psychology***	***CES-D>16***	***CES-D<16***
Male	120	68	54	134
Female	8	20	10	18
Total	12	88	64	152

Fig. 1. Distibution of the research done [7]

A similar problem was undertaken by M. de Choudhury, M. Gamon, S. Counts, and E. Horvitz [9], but their study has been vastly outdated due to their old network architectures and more involved method of collecting data. M. Park, C. Cha, and M. Cha, in 2012 explored how Twitter can be used to capture real-time moods from its users [10]. In their next work, they conducted face-to-face interviews with a selected control group of active Twitter users to discuss their depressive habits.

Another recent research done by R. Xu and Q. Zhang, in 2016 tried to gain insight into how online users approach problems related to depression from the viewpoint of social networks and linguistic trends [11]. Another more recent study by G. Shen et al. also tried to detect MDD using a more novel approach of a multimodal dictionary that was created after analysis of the data [12].

Guntuku et al. [13] conducted a very insightful review of all the previous work done on this topic, and the results were quite underwhelming. They finally concluded by recommending utilizing these automated methods to calculate a risk factor and then augment it with existing screening procedures.

All these previous works conducted their experimentation on a tiny and often controlled sample of participants.

III. METHODOLOGY

We train various models like Decision Trees, Logistic Regression, Naive Bayes, and Long Short-Term Memory (LSTM) models to varying success. This approach is then used to classify and differentiate a set of random tweets afterward to get the accuracy, Precision, recall, F1 Score, and also various other metrics.

A. *Dataset*

As no such dataset exists presently, that is in the open domain, a dataset for this task was mined and created.

The Neutral Set was formed by a set of random tweets from the Kaggle Dataset "Sentiment140". It consists of around 1.6 million tweets and is used for sentiment analysis. Three hundred thousand tweets were then chosen from this set to act as the non-depressed set of tweets.

The second part of the dataset had to be mined. Snscrape tool was used to scrape around 30000 tweets, which contained a keyword list that explicitly was created for this. The keyword list contained several hashtags and keywords that are very closely related to depression, like the name of anti-depressants, keywords like self-harm and loathing, and various other items too.

We would have opted to use the CLPSYCH dataset for our work, but it is not yet available to the public as of this time. We hope to revisit this later and test on a more accurate dataset.

B. *Data Preprocessing*

Data preprocessing was carried out in various steps

- The various tweets manually scraped were first loaded and combined into another Dataframe holding the depressed set of tweets
- The columns were filtered to only include the tweet contents and then passed through the package FTFY to fix the broken ANSI and Unicode characters in the corpus.
- Empty rows were then removed, and the corrupted rows cleaned up.
- We then saved this set of 32356 tweets to a final file as a final cleaned set of "Depressed" tweets.
- For the neutral set of tweets, we loaded the sentiment140 dataset, and then stipped the target labels.
- The text was then segregated and cleaned using FTFY to fix encoding.
- This set was saved as a CSV containing all our neutral tweets.
- We shuffled and loaded both the sets and added them to one final Dataframe.
- The data was also processed to expand all contractions.
- We also proceeded to turn the data into lowercase and removed extra spaces.
- We also stemmed the words.

- Finally, the dataset was split into test and train sets with 5% data in the test set and the rest 95% in the train part of the dataset.

C. *Data Exploration*

All the models used a bag of words approach for processing the text. The approach was very apt as after extensive data analysis and word frequency examination, it was evident that the data can be classified solely based on the word composition.

The corpus had broken Unicode, which needed to be fixed. Additionally, the usage of hashtags and links also made the data unclean. The fixing of the Unicode items was done by the use of a python module named FTFY, which specializes in fixing ANSI and Unicode format discrepancies.

The other part of the cleanup was more involved, but the issue was fixed using a module. The python module called tweet-preprocessing was used to filter out the mentions and hashtags in the text. It was also used to clean up unrelated links along with any Twitter specific tweet quotations.

A set of Stop-words was made from the NLTK module and then removed them before analyzing the the dataset.

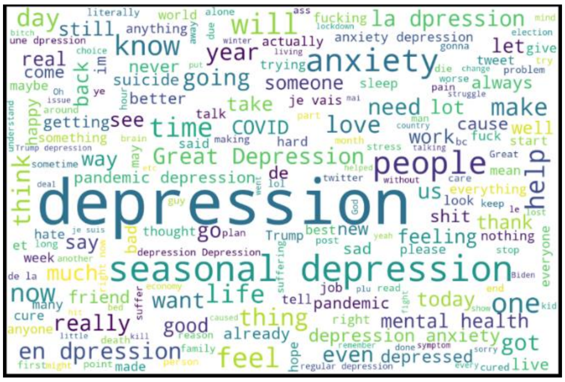

Fig. 2. WordCloud of depressed users

The Wordcloud suggests that these words were very frequently used by the control group of “depressed” users. This gives us an idea in what to expect, some items like “COVID” were unexpected, but it does seem like this is solely because of current events.

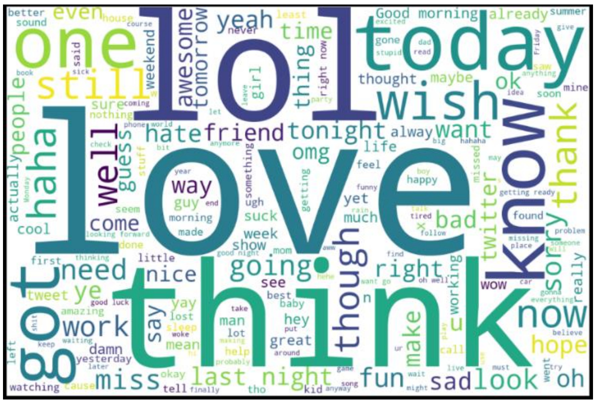

Fig. 3. WordCloud for normal users

We can clearly see the stark contrast between both the clouds; the range of words is also entirely different for the different groups.

Continuing to more in-depth frequency analysis of the data

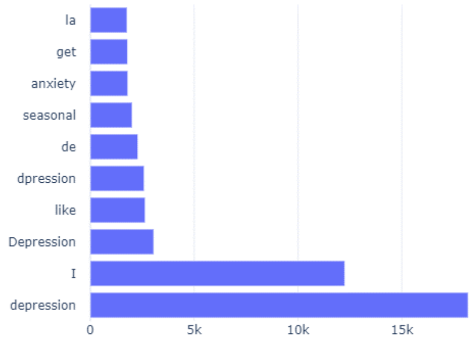

Fig. 4. Word Frequency in depressed users

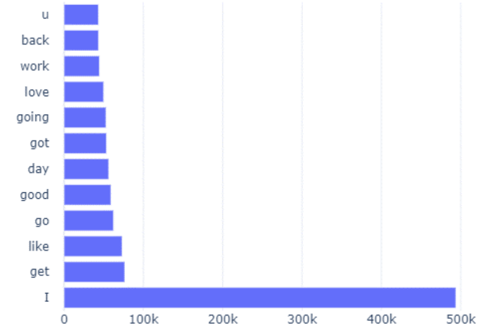

Fig. 5. Word Frequency in normal users

Therefore we can conclude that the word frequency and even the vocabulary of both the sets of users are vastly different.

D. *Model Architecture*

We made different models and compared their scores together. The bag of words approach converts the training corpus into a set. I in the sample, the frequency of each word is counted; this is then used as a feature for the different models we tested.

We kept a constant dictionary of the top 10000 words for all our computations.

1) *Naive Bayes:*

Naive Bayes classifiers belong to the family of “probabilistic classifiers”. They are based on applying Bayes’ theorem with an assumption of healthy independence between the features. They are highly scalable and work well with more simplistic feature extraction techniques like the bag of words method we are using in this case.

The probability of a given class y is given as:

$$P(\, y \mid x_1, \dots, x_n \,) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1,\dots,x_n)} \qquad (1)$$

Where x_1 through x_n are feature vectors. Keeping in mind that there is a conditional independence

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y) \qquad (2)$$

2) *Logistic Regression:*

The next model that we used was logistic regression, which, despite the name, is a linear model for classification. It is also known as maximum entropy classification, log-linear classification, or logit regression. Here we apply l_2 regularization and use “sag” solver.

The regularization problem consists of minimizing the following cost function:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log\left(\exp\left(-y_i(X_i^T w + c)\right) + 1\right) \qquad (3)$$

The “sag” solver uses Stochastic Average Gradient descent. Due to the size of our dataset the “sag” solver proved to be the best choice.

3) *Random Forest Classifier*

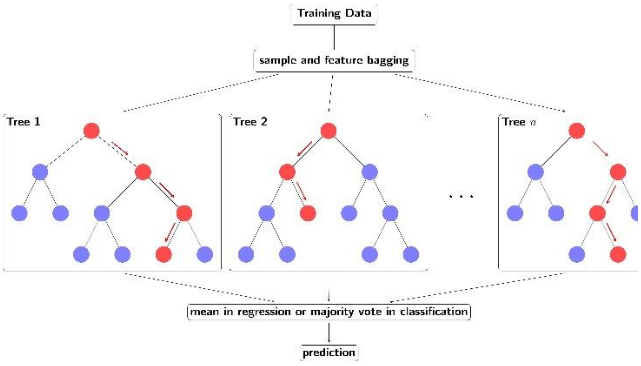
Training Data
sample and feature bagging
Tree 1
Tree 2
Tree n
. . .
mean in regression or majority vote in classification
prediction

Fig. 6. Illustration of a Random Forest Classifier [14]

The last machine learning algorithm that we tested was the Random Forest Classifier. It is an estimator that constructs many decision trees as a form of a “forest” and then uses averaging to improve the accuracy and combat overfitting.

We used the default estimator limit of 100 for our model. It used “Gini” index function to measure the quality of the splits for the different trees it constructs. The trees are also auto pruned before the final averaging.

4) *Long Short Term Memory (LSTM) Model*

LSTM networks are a prevalent form of neural networks, they are classified under the Recurrent Neural Network (RNN) category of networks. Unlike normal RNNs that use a simplistic way of recurrence i.e. a central set amount of memory that the model can consider before predicting. LSTMs can automatically find which items require more memory for predictions and hence can have a “longer” memory (hence the name LONG short term memory). This is just another weight that will be trained for different features as the model trains. While these models are an excellent choice for NLP, better model architectures like transformer networks and attention networks have since superseded the once prevalent LSTM and GRU networks.

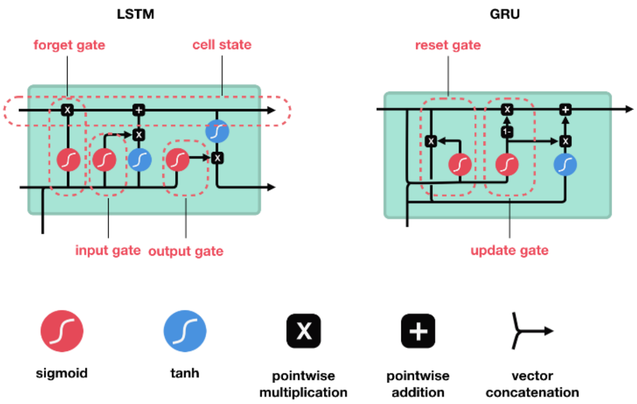

Fig. 7. Illustration of LSTM and GRU networks [15]

The models use “Gates” to control the flow of the features and have weights and biases that correspond to each gate.

GRUs also work on almost the same principle but the function of the gates is different.

These gates also have an associated activation function of either tanh or sigmoid. The biases and weights get trained as the model is introduced to more parts of the corpus.

IV. Experimental Results

The dataset is was trained and then evaluated via the hold-out technique. We calculated the accuracy, F1-Score, Precision, recall, and the Area Under the Curve (AUC) of the models.

Accuracy is the most used metric and is simple to understand. It is defined as the ratio of correct predictions and the total predictions in a sample. Sometimes, it doesn’t entirely convey the correctness of the model and is therefore not a very realistic measure.

Precision is defined as the ratio of correctly done positive value observations.

$$Precison = \frac{TP}{TP+FP} \qquad (4)$$

Recall or sensitivity is defined as the ratio of correctly identified True samples from the predicted items.

$$Recall = \frac{TP}{TP+FN} \qquad (5)$$

Further, the F1 Score is defined as the weighted average of the measures recall and Precision. It is given by

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \qquad (6)$$

The results across the board have been immensely better than the old studies done by a margin of 10% or more.

ROC curves are plots of the True Positive Rate and False Positive Rates of a model. It is a much better translation of the correctness of the model.

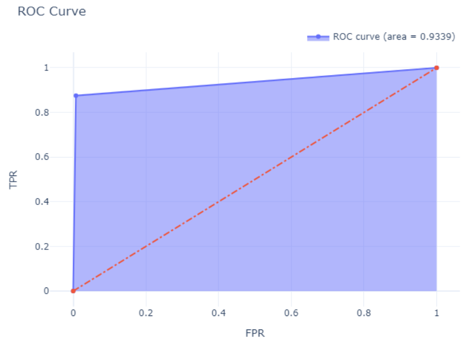

Fig. 8. ROC curve for the Naïve Bayes model

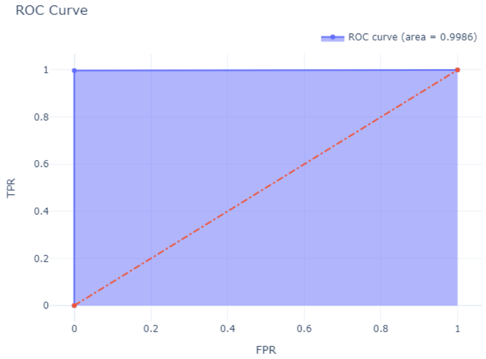

Fig. 9. ROC curve for the logistic regression model

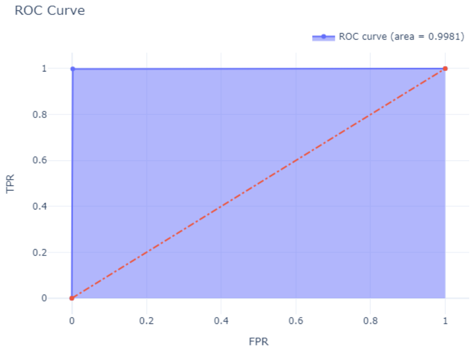

Fig. 10. ROC curve for Random Forest model

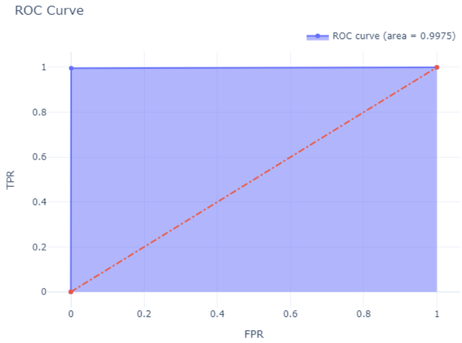

Fig. 11. ROC curve for the LSTM model

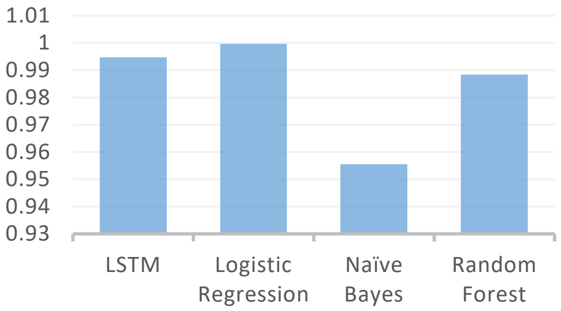

Fig. 12. ROC Score comparison for all the proposed techniques

As seen from the scores, all the models are performing really well with the accuracies in the high 95%. We think this is due to the usage of the bag of words approach, which helps to generalize the model really well among all the models and makes the True class of the tweet really easy to find. Further, the other measures of performance also show a similar view.

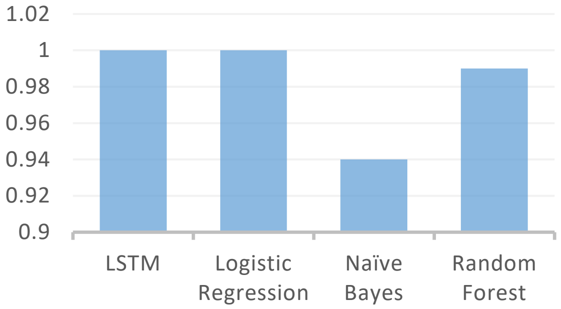

Fig. 13. F1-Scores comparison for all the proposed techniques

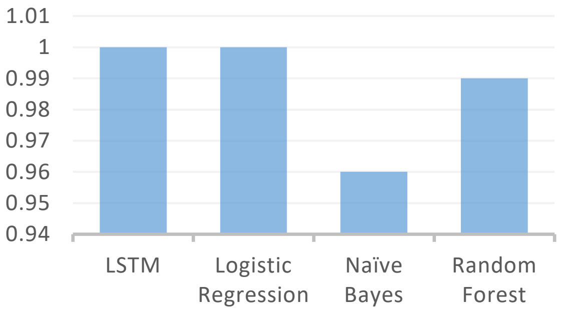

Fig. 14. Comparison of Precision for all the proposed techniques

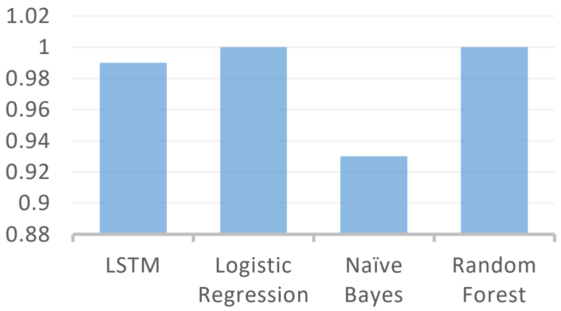

Fig. 15. Comparison of Recall for all the proposed techniques

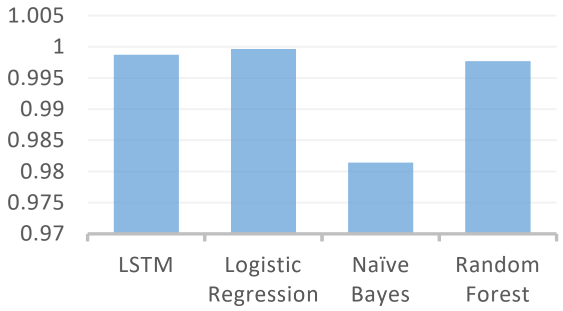

Fig. 16. Comparison of Accuracy of all the proposed techniques

Looking at all these scores, we find that surprisingly the Logistic regression model works the best, but it's all in the margin of error as the scores only differ by less than 0.01 between Random Forest, Logistic Regression, and the LSTM models.

V. Conclusion

These models were coded in python 3.8. The Naïve Bayes model was our proposed technique, other models were implemented with the help of Scikit and TensorFlow.

We were able to achieve better performance than the previous research that were performed, but the dataset might be too simplistic. Additionally, the detection of trolling and sarcasm on the internet is a significant fallback of current models as they can't encapsulate and capture sarcasm on the internet in an efficient manner. Due to the current state of internet culture, sarcasm is an integral part of it.

A review of previous studies also points to the future direction of this topic. It will be quintessential in helping and curbing depression, especially now as more and more people are spending time alone. With further advancements in the area of Natural Language Processing, large scale screening of social media data might become a real prospect.

The ethical and privacy concerns brought up are a real factor in these developments as doing this without the user's knowledge is a morally grey area.

VI. Future Work

We hope to revisit this issue with the advent of a better dataset and better deep learning models that help us to understand the language more—hopefully tackling the problem of sarcasm. We also hope to explore more types of models and also the usage of state-of-the-art Attention and Transformer models like BERT, ELMO, and XLNet.

This can then be easily implemented on a compatible social media platform, and the users assigned a risk score. Higher risk score users can then be prompted to do a more normal screening to help them identify their issues.

Acknowledgments

I want to thank all my colleagues who helped me study this topic and conduct this research. I would like to thank my mentor and teacher Dr. Ruchika Malhotra for her guidance and help throughout this research. I would also like to thank Delhi Technological University, which helped me get access to the research papers and various other resources.

References

[1] "The world health report 2001 — Mental health: new understanding, new hope.," Bulletin of the World Health Organization, vol. 79, no. 11, 2001, doi: 10.1590/S0042-96862001001100014.

[2] R. Detels and C. Chuan Tan, "The scope and concerns of public health," in Oxford Textbook of Global Public Health, 2015.

[3] C. R. Cloninger, D. M. Svrakic, and T. R. Przybeck, "Can personality assessment predict future depression? A twelve-month follow-up of 631 subjects," Journal of Affective Disorders, vol. 92, no. 1, 2006, doi: 10.1016/j.jad.2005.12.034.

[4] A. T. Beck, R. A. Steer, and G. K. Brown, "Manual for the Beck depression inventory-II," San Antonio, TX: Psychological Corporation, 1996.

[5] A. M. Abdel-Khalek, "Can somatic symptoms predict depression?," Social Behavior and Personality, vol. 32, no. 7, 2004, doi: 10.2224/sbp.2004.32.7.657.

[6] T. E. Oxman, S. D. Rosenberg, and G. J. Tucker, "The language of paranoia," American Journal of Psychiatry, vol. 139, no. 3, 1982, doi: 10.1176/ajp.139.3.275.

[7] S. Chellappan and R. Kotikalapudi, "Associating Depressive Symptoms in College Students with Internet Usage Using Real Internet Data," IEEE Technology and Society, 2012.

[8] M. A. Moreno et al., "Feeling bad on facebook: Depression disclosures by college students on a social networking site," Depression and Anxiety, vol. 28, no. 6, 2011, doi: 10.1002/da.20805.

[9] M. de Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," 2013.

[10] M. Park, C. Cha, and M. Cha, "Depressive moods of users portrayed in Twitter," in Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD), 2012, vol. 2012.

[11] R. Xu and Q. Zhang, "Understanding online health groups for depression: Social network and linguistic perspectives," Journal of Medical Internet Research, vol. 18, no. 3. 2016, doi: 10.2196/jmir.5042.

[12] G. Shen et al., "Depression detection via harvesting social media: A multimodal dictionary learning solution," in IJCAI International Joint Conference on Artificial Intelligence, 2017, vol. 0, doi: 10.24963/ijcai.2017/536.

[13] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," Current Opinion in Behavioral Sciences, vol. 18. 2017, doi: 10.1016/j.cobeha.2017.07.005.

[14] https://tex.stackexchange.com/questions/503883/illustrating-the-random-forest-algorithm-in-tikz

[15] https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21