Myriad Machine Learning Approaches for Stress Detection

Dr. Satvika[1] and Dr. Akhil Kaushik[2*]

[1]*Assistant Professor, TIT&S Bhiwani, India*
[2]*Assistant Professor, TIT&S Bhiwani, India*

Abstract: ***The modern man uses the social media platforms for multiple purposes like gaining knowledge, listening to music, watching videos or reels but the most common intent is to express his feelings towards any person, event or entity. These social media posts are thus stated as a reflection of one's feelings and can be analyzed for detecting stress in his/ her life. This paper discusses one such approach based on the various machine learning classifiers, out of which Adaptive Boosting algorithm produces best output.***

Keywords: **Sentiment Analysis, Machine Learning, Support Vector Machine, Naïve Bayes, Natural Language Processing, Social Media, Opinion Mining.**

I. INTRODUCTION

In current era, where the Web is overburdened with a large amount of text data, sentiment analysis is a highly debated and fascinating topic. Although the data is available in various formats like text, image, video, audio, animation, etc., but textual data is the most crucial one. This textual data is packed with valuable information that, when examined in light of specific criteria, can be quite useful. Finding or identifying the positive, negative, or neutral ideas, views, attitudes, impressions, emotions, and feelings expressed in the text is what sentimental analysis, also known as opinion mining, entails [1]. It has revolutionized the way to understand people or exactly existing customers, prospect customers, anxious citizens, future voters, etc. It is the upcoming domain that attracts all the major businesses and governments to better comprehend about the people that matter most to them. Opinion mining takes advantage of Natural Language Processing or popularly known by its acronym NLP, which can be defined as the branch of linguistics and artificial intelligence that focuses on teaching computers to comprehend sentences or words in spoken or written human language [2]. It was created to simplify users' lives and facilitate easy natural language communication with the system. Nowadays, plethora of netizens write reviews about almost everything on the planet and this data from internet reviews can be used to gather business intelligence by utilizing natural language processing. This huge pile of data especially the textual data can be downloaded from any social networking platforms and exploited in myriad ways for sentiment analysis. For example: User reviews about COVID vaccines can be extracted from Twitter platform and then analyzed the people's emotions about them [3], as shown in figure 1.1. Hence, it can be said that natural language processing (NLP) is crucial in extracting useful information that can aid in decision making, administration reporting, and research since a significant quantity of relevant information is embedded in the unstructured data.

Now, the web or especially social media is not limited to the teenagers or young people but has outreached to the children and older generations. As more and more people are spending enhanced time on the internet, citizens are becoming netizens. Due to this trend, their social life is getting disturbed and stress has become a normal phenomenon. People's mental health is being jeopardized by stress, anxiety, and depression. Everyone has a reason for being stressed out. People frequently share their emotions on social media platforms such as Instagram in the form of

posts and stories, and on Reddit in the form of subreddits where they ask for suggestions about their lives. Many content creators have stepped forward in recent years to create powerful content to assist people with their mental health [4]. Many organizations can use stress detection to identify which social media users are stressed so that they can help them as soon as possible. Stress detection is a daunting task because there is a vast collection of words that mundane persons can use in their posts to indicate whether he/she is experiencing psychological pressure.

Figure 1. Twitter as a Tool for Sentiment Analysis

II. RELATED WORK

A. Hogenboom et. al., suggested that the people are increasingly use emoticons in the social media posts to express, emphasize, or disambiguate their sentiment, and hence it is critical for automated sentiment analysis tools to correctly account for such graphical cues for sentiment. In this work, the researchers have taken the tweets in Dutch language and also data extracted from forums, that contained manually annotated emoticons. The experiments revealed that the emoticons prove to be a potent proxy for detecting user sentiments [5].

S. Yoon et. al., proposed that the use of sentiment analysis, natural language processing, and visualization techniques provides research teams with insights into large volumes of daily self-report stress notes taken from smartphones. Positive emotion scores derived from qualitative data by three machine learning algorithms provide quantified descriptive contextual information on low level self-rated stress scores [6].

M. J. Lee, et. al., discussed sentiment analysis of social media data was used to investigate the impact of the Sewol Ferry disaster on social stress. In this study, data from YouTube, Twitter, and Facebook were used. ANOVA was used to compare negative, neutral, and positive sentiments with a 95% confidence level. The results of the study revealed a significantly negative sentiment across all social media platforms [7].

E. Tromp & M. Pechenizkiy suggested – SentiCorr, which is an automated sentiment analysis system for multilingual user generated content from social media and e-mails. Unlike most existing systems that use sentiment classification and opinion mining to answer marketing questions, this proposed system is designed to

help the individual users become more aware of the sentiment in their correspondence [8]. Working of this system is demonstrated in the figure 1.2 below.

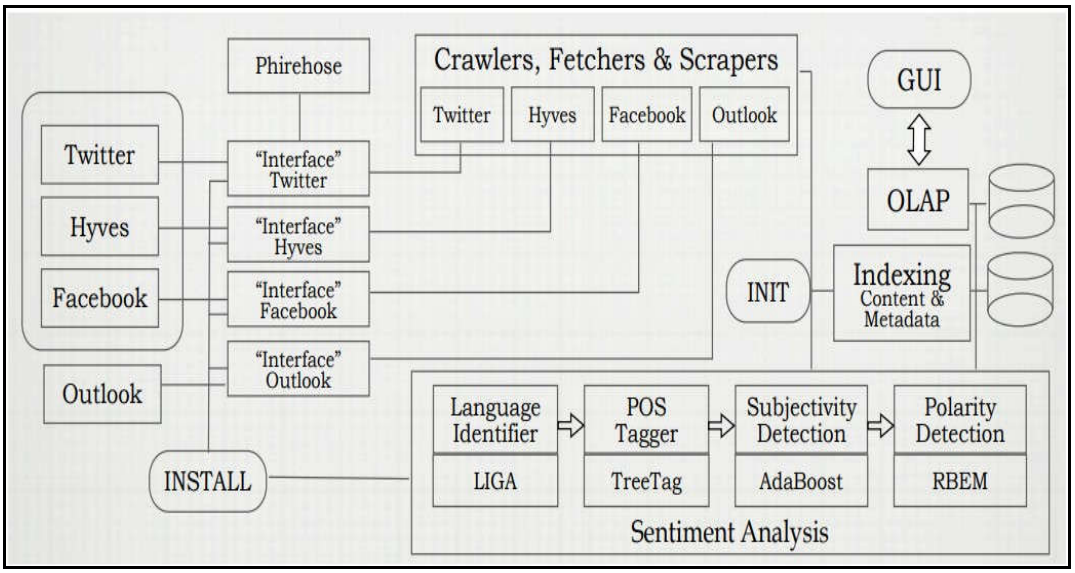

Figure 2. Overview of SentiCorr Framework [8]

L. Stappen introduces Multimodal Sentiment Analysis (MuSe) is a challenge that focuses on sentiment and emotion recognition, as well as physiological emotion and emotion-based stress recognition, by more thoroughly integrating audio-visual, language, and biological signal modalities. The purported system use the MuSe-CaR dataset for user-generated reviews and introduce the Ulm-TSST dataset for people in stressful situations [9].

III. PROPOSED SYSTEM

This section illustrates how to use machine learning techniques in conjunction with feature extraction procedures to perform sentiment analysis on textual data. The data that is used for prediction or classification analysis can be taken from several sources like data repositories like Kaggle, Zenodo, etc., or extracted from social media platforms like Facebook, Twitter, Reddit, Snapchat, Instagram, etc., or from other commercial platforms like Amazon or Flipkart. The dataset used in this study is taken from Kaggle data repository.

The data is then pre-processed, which includes removing hashtags, hyperlinks, and special symbols such as @, etc. from the tweets. Additionally, words from languages other than English are removed, and tokenization is performed as a result. Finally, all stopwords (frequent words with little relevance) are removed, followed by token stemming or lemmatization, etc. This process converts the words to their root form, thus removing ambiguity in the natural languages [10]. Also, the removal of duplicates is done, so that the data is ready for sentiment classification.

The next step is feature extraction, which is accomplished using an ensemble technique involving POS tagging, dependency parsing and TF-IDF. Subsequently, various machine learning algorithms like Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, Random Forest, Decision Tree, etc. are applied on the cleaned data for doing the sentiment analysis. Apart from the basic machine learning algorithms, advanced techniques like Adaptive Boosting is also employed. Then using sci-kit learn library, the data is split into 80-20 ratio, where 80% of data is

used for training and the remaining 20% data is utilized for testing objectives. This whole process is illustrated in the figure 1.3 below.

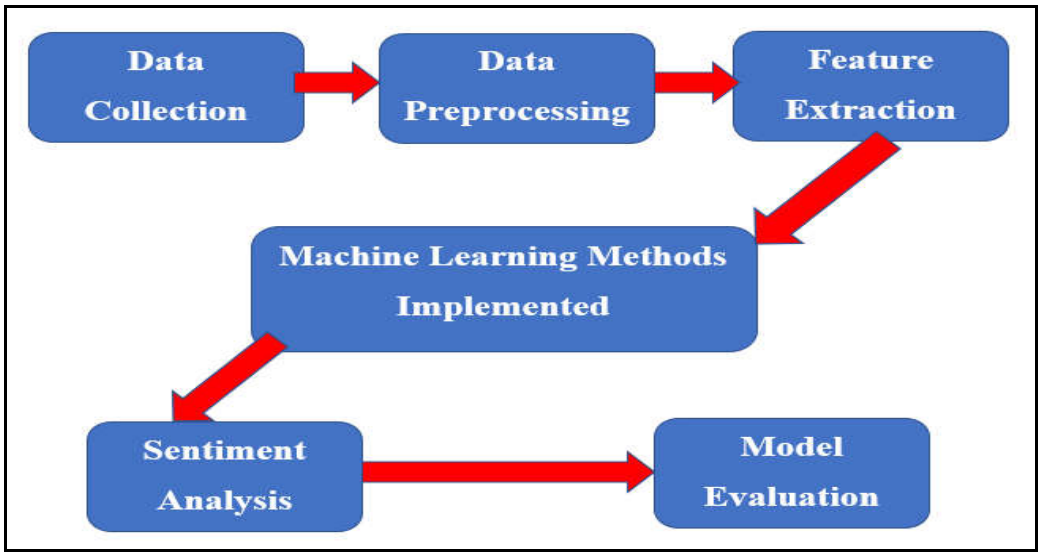

Figure 3. Proposed Methodology

IV. RESULTS AND DISCUSSION

The Stress dataset used in this research is taken from the Kaggle data repository. It contains 2838 rows and 116 columns, although only the columns of utmost importance are "text" and "label" that represent the input sentence and corresponding sentiment associated with it. There can be numerous words that are frequently used in the collected data and they can be easily traced out using the wordcloud. The following wordcloud (in the outline of African continent) displays the most frequent terms used by the users in the reddit posts.

Figure 4. Wordcloud showing repeated terms

For improved user comprehension, the sentiment distribution among the data i.e., sentences reflecting stress vs sentence displaying no stress is checked and visualized, as shown in figure 5 underneath.

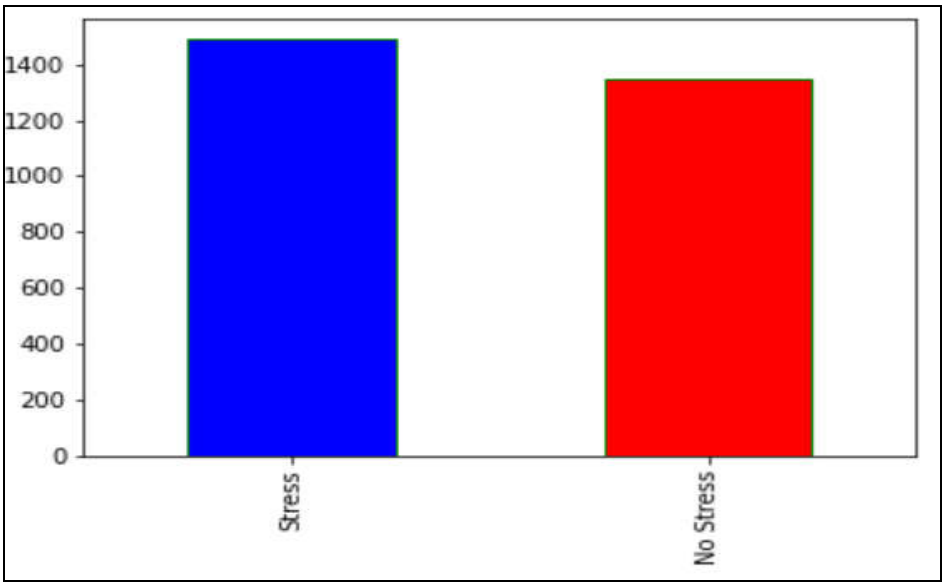

Figure 5. Sentiment Distribution in the Dataset

After the data cleansing and division, the machine learning algorithms are applied for classification on the data one by one and their performance metrices is recorded. The machine learning algorithms and their values for accuracy, precision, recall and f1-scores is listed in the following table 1.

Table 1. Performance Metrices of Various Machine Learning Algorithms

Models	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	74.1%	86.46%	88.43%	86.78%
Support Vector Machine	73.9%	82.67%	85.38%	83.71%
K-Nearest Neighbour	71.5%	78.42%	79.11%	81.58%
Random Forest	73.5%	79.91%	70.37%	72.63%
Decision Tree	63.5%	75.38%	66.37%	68.26%
Adaptive Boosting	81.8%	89.32%	87.75%	89.94%

As illustrated in the table above, the performance of Adaptive Boosting is better than the Naïve Bayes, Support Vector Machine, K-Nearest Neighbour, Random Forest, Decision Tree classification algorithms.

V. CONCLUSION

This paper discusses how a stress detection model can be created, which can input any sentence or user post taken from any social networking platform and predict whether the user is experiencing any stressful conditions in his life currently or not. This model can thus be utilized to analyze user posts from social media and provide helpful insights about the stress in any person's life. This model can be of great help to NGOs or even MNCs who want to provide a healthy working environment to its employees. Several machine learning algorithms were employed on the dataset taken from Kaggle and Adaptive Boosting algorithm displays best output when compared to its counterparts like NB, SVM, K-NN, Decision Trees and Random Forest. Although, the model efficiency can be enhanced profusely by providing more training data.

REFERENCES

[1] *A. M. Mohsen, A. M. Idrees, and H. A. Hassan, "Emotion analysis for opinion mining from text: a comparative study", International Journal of E-Collaboration (IJeC), vol. 15, no. 1,* ***(2019)****, pp. 38-58.*

[2] *W. Andrew, S. Fu, S. Moon, Md. El Wazir, A. Rosenbaum, V. C. Kaggal, S. Liu, S. Sohn, H. Liu, and J. Fan, "Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation", NPJ digital medicine, vol. 2, no. 1,* ***(2019)****, pp. 1-7.*

[3] *M. T. J. Ansari, and N. A. Khan, "Worldwide COVID-19 Vaccines Sentiment Analysis Through Twitter Content", Electronic Journal of General Medicine, vol. 18, no. 6,* ***(2021)****.*

[4] *B. Stahl, and E. Goldstein, "A Mindfulness-Based Stress Reduction Workbook", New Harbinger Publications,* ***(2019)****.*

[5] *A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak, "Exploiting emoticons in sentiment analysis", Proceedings of the 28th annual ACM symposium on applied computing, pp. 703-710,* ***(2013)*** *March 2013.*

[6] *S. Yoon, F. Parsons, K. Sundquist, J. Julian, J. E. Schwartz, M. M.Burg, , ... and K. M. Diaz, "Comparison of different algorithms for sentiment analysis: psychological stress notes", Studies in Health Technology and Informatics, vol. 245, 1292,* ***(2017)****.*

[7] *M. J. Lee, T. R. Lee, S. J. Lee, J. S. Jang, and E. J. Kim, "Machine learning-based data mining method for sentiment analysis of the Sewol Ferry disaster's effect on social stress", Frontiers in Psychiatry, 11, 505673,* ***(2020)****.*

[8] *E. Tromp, and M. Pechenizkiy, "Senticorr: Multilingual Sentiment Analysis of Personal Correspondence", Proceedings 2011 IEEE 11th International Conference on Data Mining Workshops, pp. 1247-1250,* ***(2011)*** *December 2011.*

[9] *L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E. M. Messner, ... and B. W. Schuller, "The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion, Physiological-Emotion, and Stress", Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge, pp. 5-14,* ***(2021)****.*

[10] *S. García, J. Luengo, and F. Herrera, "Data Preprocessing in Data Mining", Cham, Switzerland: Springer International Publishing, vol. 72,* ***(2015)****, pp. 59-139.*