A Novel Ensemble Approach for Toxic Comment Detection Using Context-Free and Context-Aware Models

Sahana. P. Shankar[1]
Assistant Professor
Department of Computer Science and Engineering
Ramaiah University of Applied Sciences, Bengaluru
Karnataka, India

Pranathi Hegde[2]
Student
B. Tech in Information Science and Engineering
Ramaiah University of Applied Sciences, Bengaluru
Karnataka, India

Nidhi N P[3]
Student
B. Tech in Computer Science and Engineering
Ramaiah University of Applied Sciences, Bengaluru
Karnataka, India

Sanjana N[4]
Student
B. Tech in Information Science and Engineering
Ramaiah University of Applied Sciences, Bengaluru
Karnataka, India

Sree Bhanu Mukkamala[5]
Student
B. Tech in Information Science and Engineering
Ramaiah University of Applied Sciences, Bengaluru
Karnataka, India

Abstract:- With more than 100 million active users on social media today, it has become inevitable that the average user is exposed to some form of cyberbullying. Toxicity and hate comments have become a critical challenge necessitating efficient tools for their detection and mitigation. In this study, we propose a novel ensemble approach combining context-free and context-aware models to detect toxic comments. Using the Civil Comments dataset, we curated two distinct datasets, one with conversational context and one without, which had to be extensively processed and augmented before they were employed. The two models were built using the RoBERTa architecture which was further fine-tuned and modified to suit this particular task. Lastly, the classification outputs from both the models were integrated using equal weights. The context-free model achieved an accuracy 94.87% and an F1 score of 0.95 for both labels- toxic and non-toxic. The context-aware model showed an accuracy of 87.82% achieving an F1 score of 0.91 for non-toxic comments and 0.80 for toxic comments. This work underscores the importance of incorporating conversational context and ensemble techniques in developing robust toxicity detection systems.

Keywords:- *Social Media, Cyberbullying, Toxicity, Ensemble Approach, Civil Comments Dataset, Data Augmentation, RoBERTa.*

I. INTRODUCTION

In today's digital age, social media has become an integral part of our life. With this, there is a rapid increase in hate speech and toxicity in the online space. This affects the mental health of the user and disrupts meaningful conversations. Therefore, it is crucial to detect this toxicity to maintain a safe and inclusive online space. However, existing models fail to accurately detect this toxicity, especially when there is a case involving sarcasm, intent and contextual dependencies.

There is a huge requirement for automated solutions that can ensure politeness in online interactions. We have seen in multiple studies how important is the context to comprehend and identify toxicity, because sometimes isolated remarks do not completely convey the true meaning of the interaction. In order to have a fully functional and reliable system to detect this, it is very important to combine contextual data with sophisticated machine learning models.

In our work, two datasets from the Civil Comments Dataset are used. One with isolated comments and another with its parent comments to provide context. Class imbalance, which is a common problem in toxicity detection, is addressed by using data balancing techniques such as augmentation with GPT-Neo and downsampling. We have fine-tuned a transformer-based RoBERTa model on both datasets to detect toxicity effectively. Our approach provides higher accurate reliable results in detecting toxicity while taking conversational context into consideration by combining the predictions made from context-free and context-aware models using a weighted ensemble approach.

A weighted ensemble strategy to combine predictions from context-free and context-aware models is used in our approach, which is built on recent developments in toxicity detection. Even minor hazardous behaviours and intent with robustness is not missed by our detection system. The model built can be improvised in future, by adding multilingual datasets, multimodal data such as photos and videos, and real-time deployment for more proactive and inclusive moderation systems. As illustrated in Figure 1, the model will take input from the user and provide toxic or non-toxic classification of the provided comments baked on the context.

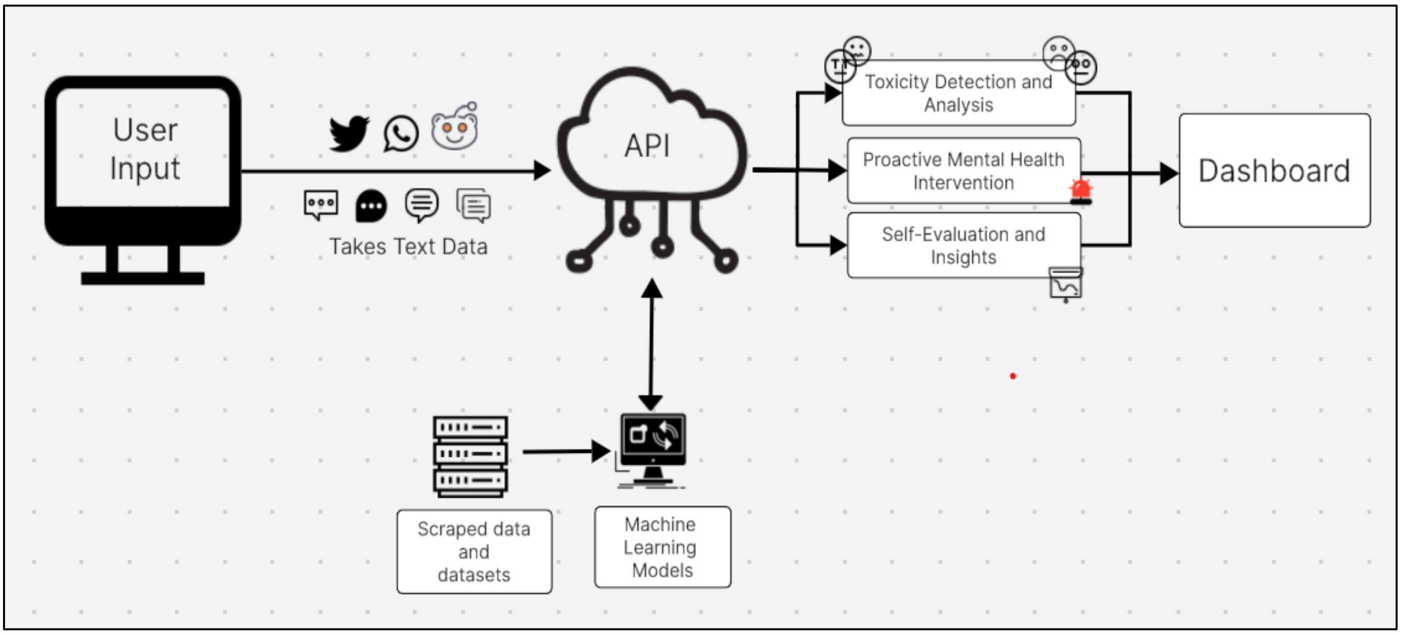
User Input
Takes Text Data
API
Toxicity Detection and Analysis
Proactive Mental Health Intervention
Self-Evaluation and Insights
Dashboard
Scraped data and datasets
Machine Learning Models

Fig 1 High-level Working of the Model

II. LITERATURE REVIEW

Due to growing use of social media platforms, emotion detection and sentiment analysis in textual data has gained plenty of attention. Effective text analysis is required because online interactions involve sarcasm, context and multilingual usage. Traditional analysis methods which depend on keyword-based analysis fail to capture emotions and contextual dependencies. In order to build a reliable system that aptly analyses text , we need to apply machine learning algorithms and deep learning techniques.

Support Vector Machine (SVM) and logistic regression for sentiment classification in Twitter dataset was done by Neethu and Rajasree (2013) [16] and Le and Nguyen (2015) [17]. Their approach focused on feature engineering and preprocessing to work with unstructured data. These research showed the limitations of traditional machine learning approaches to handle complex text structures and brought in deep learning-based approaches.

Ho et al. (2020) [2] implemented a system that has a multi-layered architecture to analyse Vietnamese text. This study showed us the potential of deep learning in low-resource languages. Batbaatar et al. (2019) [3] built a Semantic-Emotion Neural Network (SENN) to connect the understanding of semantics with neural networks, this improved the accuracy of emotion detection.

To capture syntactic and semantic nuances, Jemai et al. (2021) [18] used convolutional and recurrent neural networks on different language datasets. Many studies have shown the potential of transformer-based models in the process of toxicity detection because they manage to take care of context-dependent data.CTNet is a conversational transformer network that was made by Lian et al. (2021) [10] ,to deal with sarcasm and contextual signals. Jain and Dandannavar (2016) and Gupta et al. (2017) [22] both stressed the importance of fine-tuning transformer models to account for task-specific subtleties. When dealing with implicit sentiment and complex emotional context, these models perform better than conventional methods.

Studies of Soleymani et al. (2017) [15] show that combining textual, vista and aural data from multimodal sentiment analysis gives us a clear viewpoint and promises for improving social media analytics. Class imbalance is still a major problem here.To decrease skewed data distributions, Gaind et al. (2019) [4] and Singh et al. (2017) [24] proposed approaches such as oversampling and synthetic data augmentation. Mitra and Mohanty (2020) [20] and Gupta et al. (2017) [21] improved the classification by using ensemble techniques. Hicham et al. (2023)[8], used a similar approach to perform analysis on Arabic text.

Extensive Surveys actually help in improvising the analysis.Works of Medhat et al. (2014) [13] and Yue et al. (2019) [14] identified where the existing approaches go wrong. For example, scalability and handling sarcasm are key areas that need to be looked into. In the study of Kusal et al. (2022) [6] , many techniques and datasets were analysed. This shows how important multimodal and multilingual approaches are for real world applications. Research by Acheampong et al. (2020) [7] highlighted the importance of cross-linguistic scalability. They proposed a technique that used domain-specific features to improve generalisation.

The importance of combining manual annotations, crowd-coding, and Machine Learning (ML) algorithms is highlighted in recent studies Dang et al. (2020) [12] and Van Atteveldt et al. (2021) [23] This increases the chances to receive improved efficiency of sentiment analysis. Studies such as Wankhade et al. (2022) [11] emphasise the importance to expand datasets and to include multi-modal data, and incorporate real-time systems for wider applications.

III. METHODOLOGY

➢ *Dataset Description*

Two particular datasets have been used for our project. These datasets, which were taken from the Civil Comments Dataset, were specifically useful for examining toxic comments. Ten annotators per post initially annotated the first dataset; however, annotators were not shown the parent post, which was the thread's prior post. It includes three important fields that are focused on isolated comment classification: comment_id, comment, and toxicity_label. The toxicity_label is binary, with 0 denoting a non-toxic comment and 1 denoting a harmful comment. With this dataset, toxicity in comments may be easily assessed. For this assessment, no other contextual information was required.

Contextual data, importantly improving the approach, is included in the second dataset for a much more subtle toxic comment classification. From the first dataset 10,000 CC posts were randomly sampled and both the target and the parent post were provided to annotators who labelled the comments as toxic or non-toxic contextually. Four main fields—comment_id, comment, parent_comment, as well as toxicity_label—compose it. By including the parent_comment field, the model considers the comment's context, better understanding conversational toxicity. The toxicity_label, quite simply, is also a binary label, just like the first dataset.

However, both the datasets were imbalanced which posed considerable challenges. Compared to the number of toxic comments, there were significantly more non-toxic ones. This caused a skewed distribution which could affect the model's capacity to identify toxic remarks. Due to the skewed nature of the datasets, the majority of predictions tended to be non-toxic simply because their numbers outnumbered those of the toxic cases. In order for models to learn how to effectively identify harmful comments—even when the majority of the data was not toxic—it was necessary to rectify this imbalance.

Tables 1 and 2 show a snapshot of the two datasets used.

Table 1 Snapshot of the Dataset with Context

Id	Text	Parent	Label
5162354	Mukluk. It is often better to remain silent and have others wonder if you are a fool, than to speak and remove all doubt. You are a complete moron.	I fought fire in the 70's and also got trapped in that mess last summer on the Seward Highway for the McHugh creek fiasco. Let me say right now no one is ever going to be put at risk along that section of road ever and if APD, AFD, and the others can not figure out how to fight a fire and keep traffic flowing then I strongly suggest they look for NEW occupations. Does anyone else see an alarming trend going on here?	1
123178874.21458.21458	Read the FAR page and you will see a gusher of opinions why this page doesn't come close to deserving FAR in its current state.	No, I don't think so. Apparently the FARC was closed and the result was that the article was de-featured, but I don't see any statement of the reasons why.	0
122665847.18097.18097	If you want, I can make some changes. I'm actually a copy editor on my school paper...which may not be a huge deal in reality, but I know how to write pretty well. It'd be great for the Minneapolis article to be a FA.	Thanks for taking the time to comment. Adding {{Copyedit}} which may attract some writers, and added the article to Requests for copy editing. Also I will restore the lists I moved because now I see Elfangor801 meant the prose itself. Best wishes. -	0

Table 2 Snapshot of the Dataset without Context

Id	Text	Label
511702	Hiya, Pandora. Good to see you, too. Yes, I'm well. Thank you. My kid brother is doing well on account of his treatment. It came from The States. So God Bless America! Thank you so much, you and all the others, for your love and prayers. You're truly remarkable, the lot of you.	0
773540	And? What? A strategic plan, for what? This article said nothing.	0
6270109	Brison practises deviant sex His wife is not a she It is another deviant	1

To further examine the dataset, an analysis of the lengths of both the parent comments and the principal comments in both the datasets was carried out. As displayed in Figure 2, in the dataset with context, comments labelled as non-toxic have a longer word count as compared to toxic comments. Most comments in this dataset are concise, with the majority containing fewer than 50 words, resulting in a skewed distribution toward shorter lengths. In the dataset with context, illustrated in Figure 3 and Figure 4, due to the inclusion of the parent comments there is additional variability, however the pattern remains the same as the non-toxic comments have a longer word count as compared to the toxic comments. It was also observed that the parent comments tend to be shorter than the principal or target comment. These trends imply that toxic comments tend to be shorter than their non-toxic counterparts.

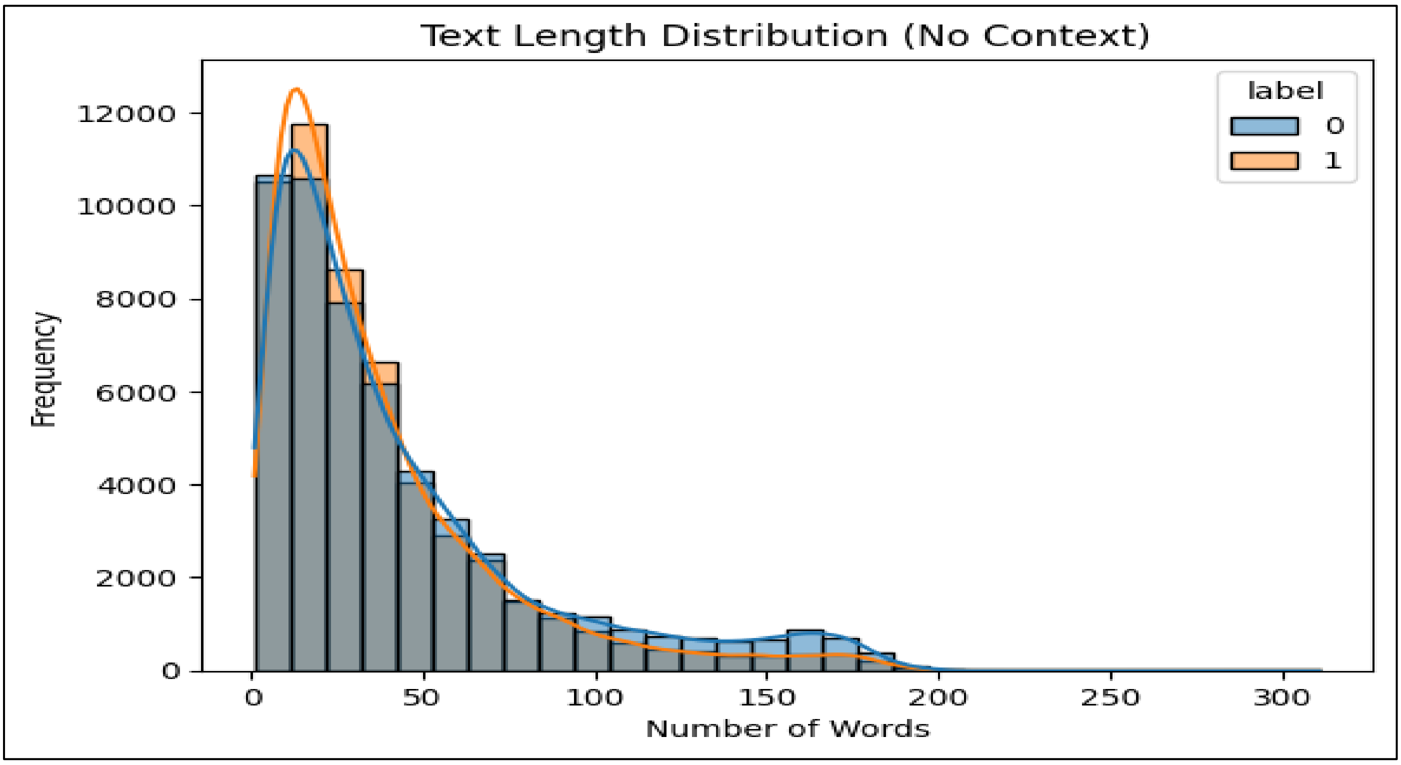

Fig 2 Text Length Distribution of Dataset without Context

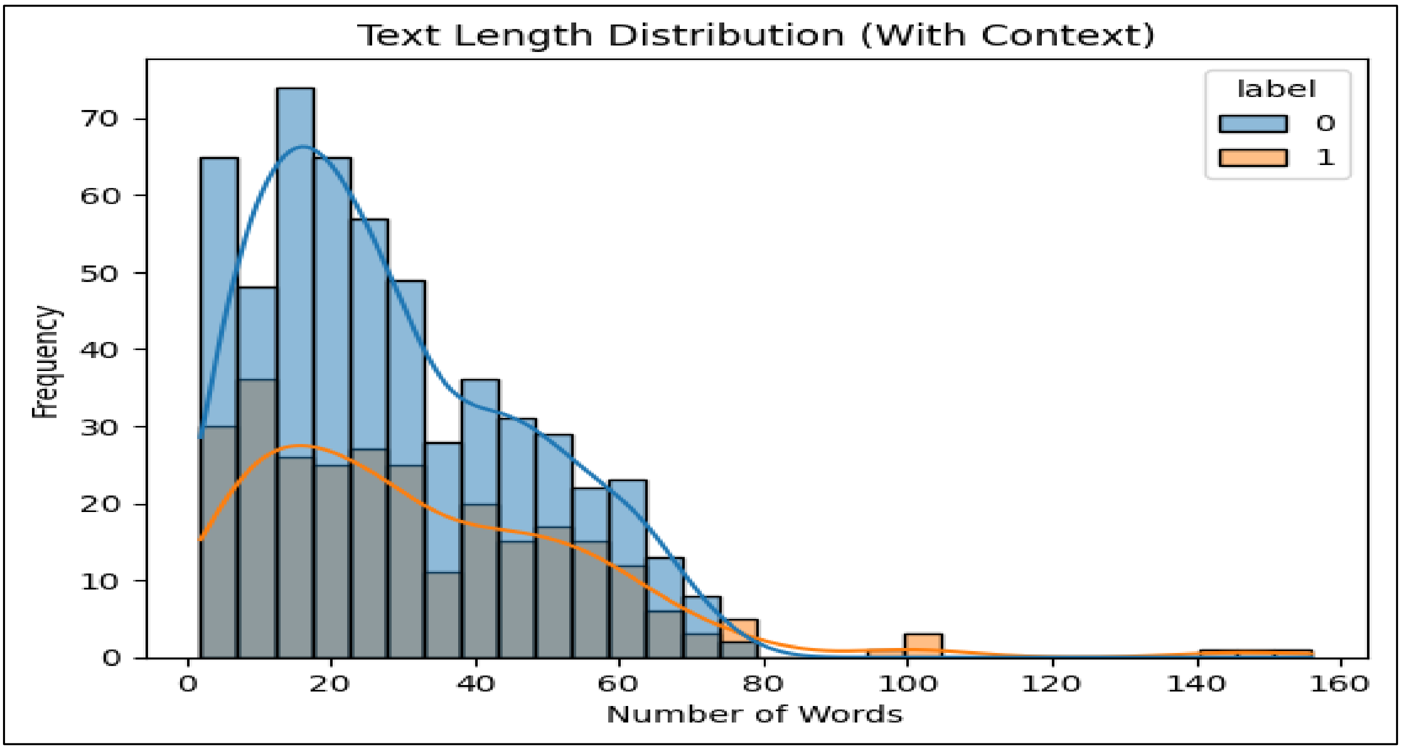

Fig 3 Text Length Distribution of Principal Comment in Dataset with Context

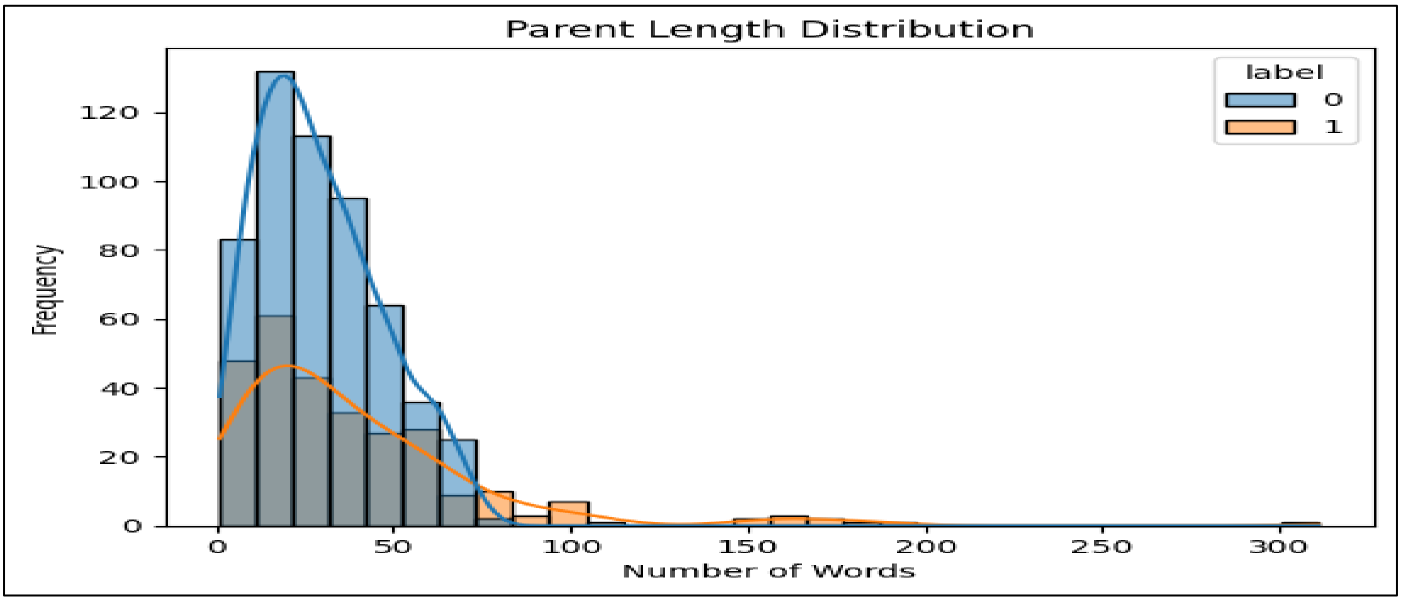

Fig 4 Text Length Distribution of Parent Comment in Dataset with Context

➢ *Data Balancing*

To ensure that the model would be robust and more accurate, several tactics were incorporated for each dataset to provide a more equal representation of both toxic and non-toxic comments. To increase the count of toxic comments several toxic remarks were added to the context-free dataset by combining multiple subsets from the Civil Comments Dataset. This process of aggregating the comments and then further downsampling the non-toxic comments yielded a consistent and balanced dataset with roughly 50,000 toxic and 50,000 non-toxic comments.

However, a different methodology was used to balance the dataset with context which included both the primary comments in the thread and their corresponding parent comments. For this purpose various methods of data augmentation were explored and compared including synonym replacement, random insertion, random deletion and character-level modifications before we honed in on the prospect of using GPT-Neo, a pre-trained language model. To generate more data instances inline with the existing instances GPT-Neo was employed to synthesize samples and artificially generate more toxic remarks using data augmentation. Our goal was to improve the quality and relevance of the additional toxic comments. In order to determine whether or not these enhanced remarks aligned with the required qualities of being poisonous, they were tested against Google Perspective AI. To balance the dataset, the non-toxic remarks were further downsampled. Together, these efforts improved the dataset's balance between harmful and non-harmful remarks, which improved the model's performance on an increasingly difficult and context-dependent task.

Figures 2 and 4 show the dataset distribution before balancing and figures 3 and 5 show the distribution after balancing.

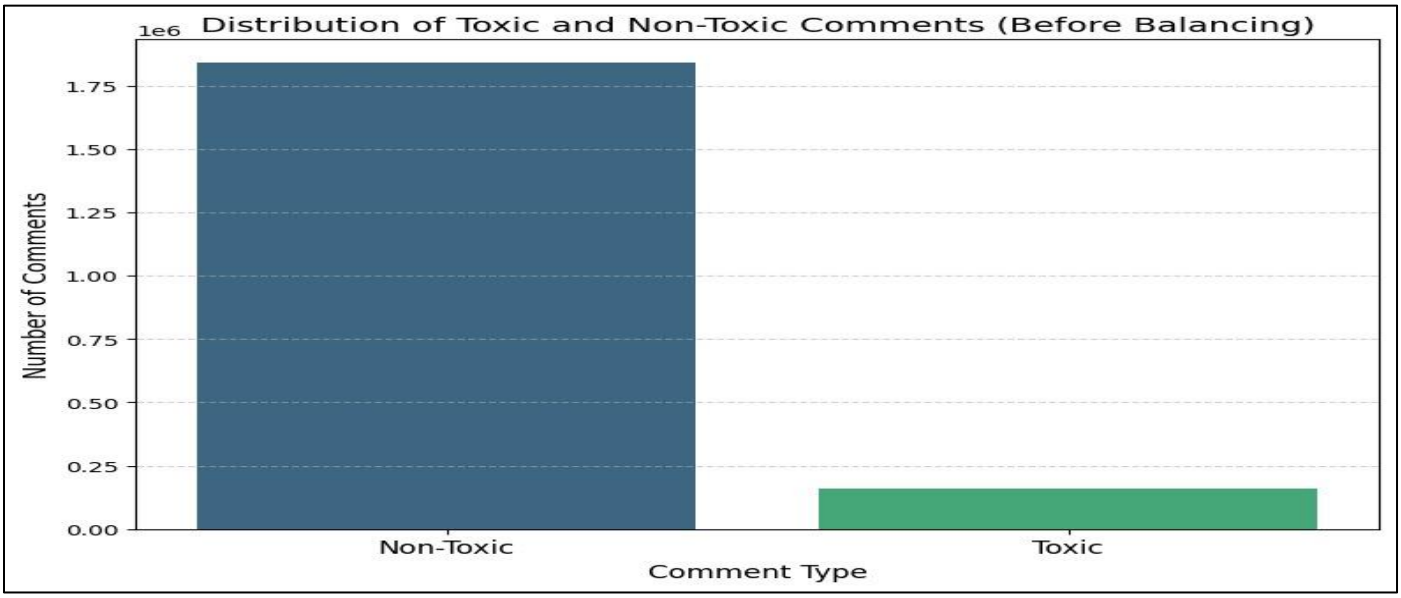

Fig 5 Dataset without Context before Balancing

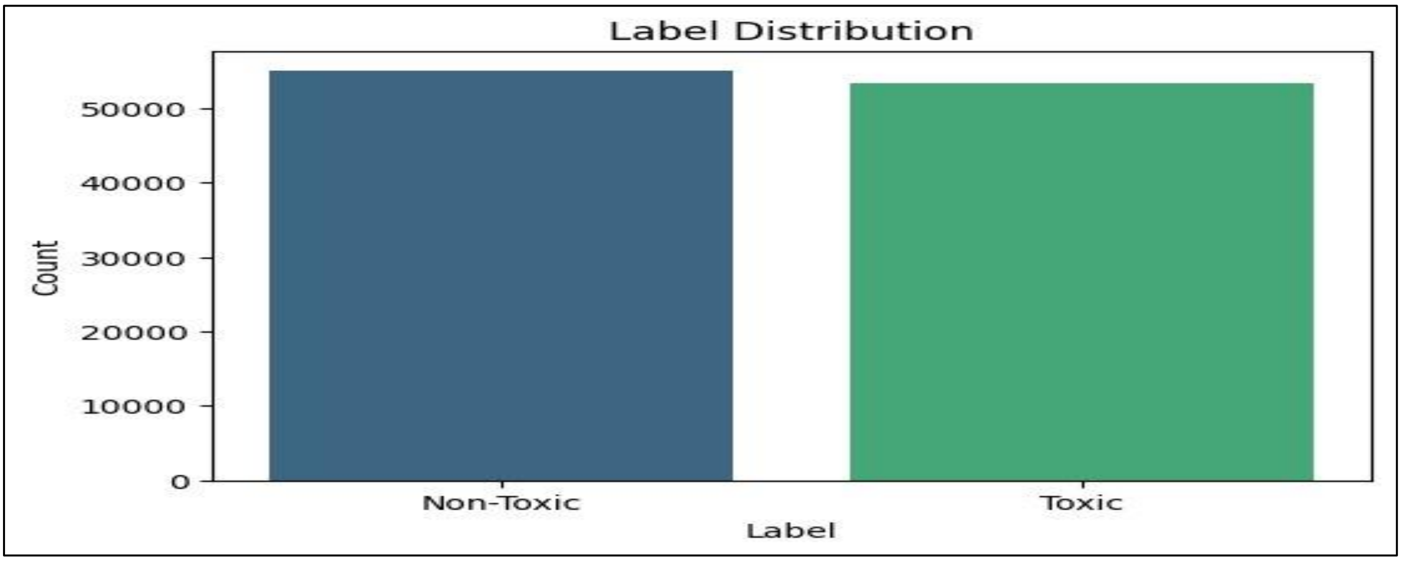

Fig 6 Dataset with Context after Balancing

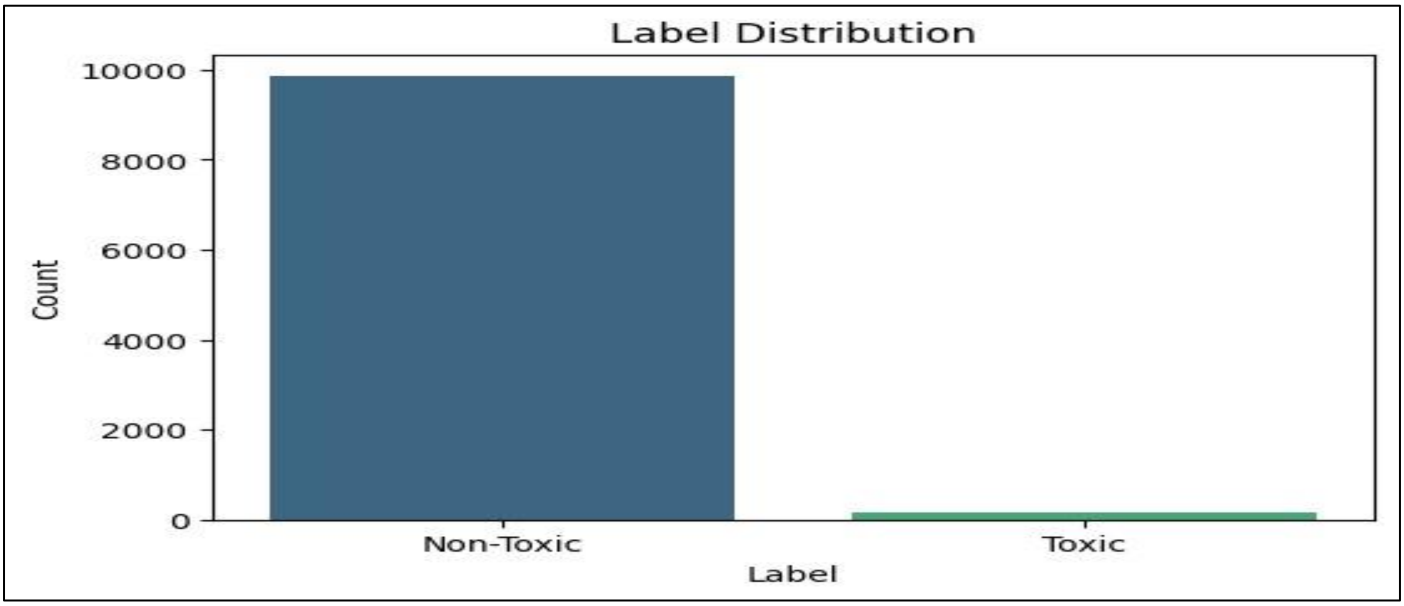

Fig 7 Dataset with Context before Data Augmentation

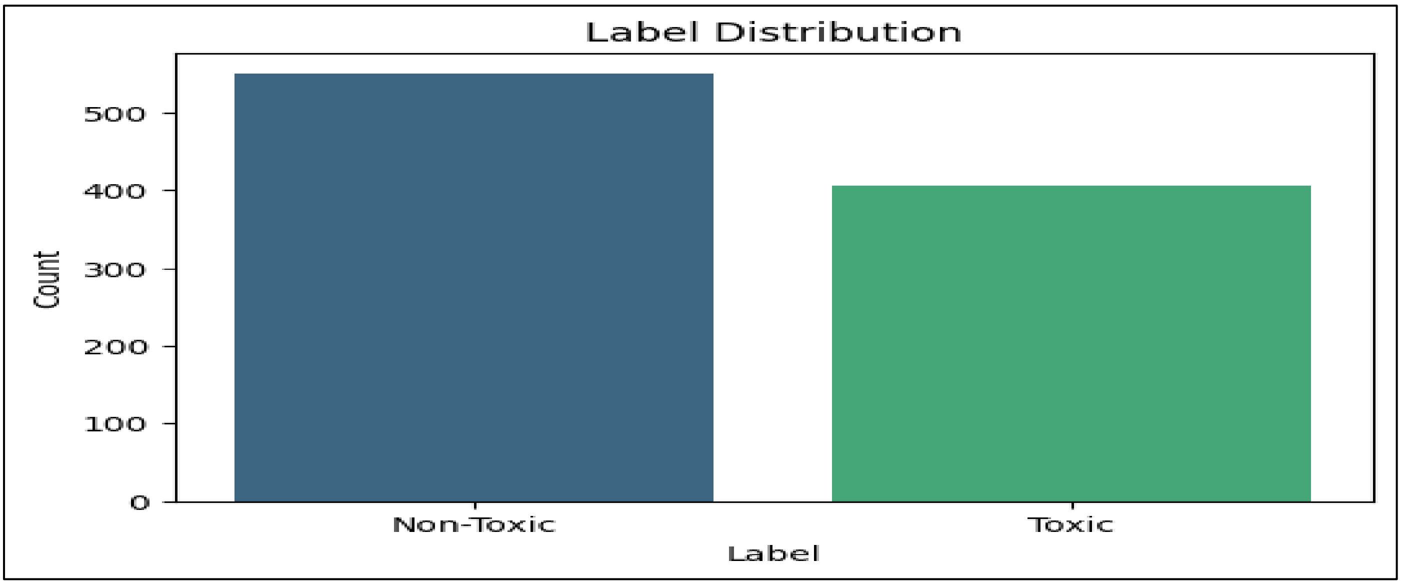

Fig 8 Dataset with Context after Data Augmentation

➢ *Classification*

A transformer-based language model, initialized with pre-trained weights from the RoBERTa architecture, served as the foundation for our classification framework. In order to capture both contextual and non-contextual components of comment classification, the model is further refined twice on two distinct datasets.

Usually, in the typical RoBERTa implementation the classification layer only uses the final [CLS] token output . This was vastly improved by extending it and modifying the last fully connected layer to match the binary classification goal (toxic or non-toxic) and adding a dropout layer to lessen overfitting. To ensure that the model's performance was enhanced on contextual and non-contextual datasets these modifications were required.

This improved model was adjusted for the context-free data using comment data and associated toxicity labels. The model will then proceed to detect harmful language and target just on comments that are independent. Simultaneously, a second instance of the updated model was trained using the context-rich dataset, integrating both the parent comment and the main comment.

Both the main comment and the parent comment were tokenized using the RoBERTa tokenizer to convert text into input IDs and attention masks. The parent comment was processed using RoBERTa to extract token embeddings which are then passed through an LSTM layer to encode the parent comment into a fixed-dimensional vector.The CLS token of the main comment and the LSTM output of the parent comment are concatenated and put through a fully connected layer to get the toxicity classification output from the with-context model. Due to this contextual approach, the framework was able to take into account conversational context, such as provocation, intent, and sarcasm.

Once classification was completed on both models, a weighted ensemble approach was applied to combine their outputs. To ensure that both perspectives were equally represented in the final decision, equal weights of 0.5 were assigned to the predictions from the context-free and context-aware models. This approach allowed for a balanced and comprehensive method of identifying and categorizing comments as toxic or non-toxic, effectively leveraging the strengths of both models.

IV. RESULTS

To comprehensively understand the effectiveness and accuracy of the two models we evaluated them against standard metrics such as precision, recall, f1-score and support. Precision is a metric that is used to calculate the proportion of correct predictions for the positive label in each case and recall helps understand how well a model identifies the positive instances in a dataset. F1-score takes the harmonic mean of the two metrics precision and recall into a single metric to denote the quality of the classifier.

The model without context showed an impressive accuracy of 94.87% while the model with context showed an accuracy of 87.82%. The lesser accuracy of the model with context can be attributed to the limited dataset. The metrics are displayed in Table1 and Table 2 for both the models.

Table 3 Metrics for Model Trained on Dataset without Context

Label	Precision	Recall	F1 Score
Non-Toxic (0)	0.97	0.93	0.95
Toxic (1)	0.93	0.97	0.95

Table 4 Metrics for Model Trained on Dataset with Context

Label	Precision	Recall	F1 Score
Non-Toxic (0)	0.85	0.98	0.91
Toxic (1)	0.95	0.69	0.80

A confusion matrix was used to evaluate the model's performance as shown in the figures below. The confusion matrices highlight the effectiveness of the model in differentiating between the toxic and non-toxic comments. For the model with context as displayed in Figure 5, out of the total 101 non-toxic comments, the model correctly predicted 96 as non-toxic, achieving a high true negative rate. For toxic comments, the model successfully identified 44 out of 55, reflecting a strong true positive rate. Further balancing of this dataset with various augmentation methods would help improve the model's ability to address the edge cases.

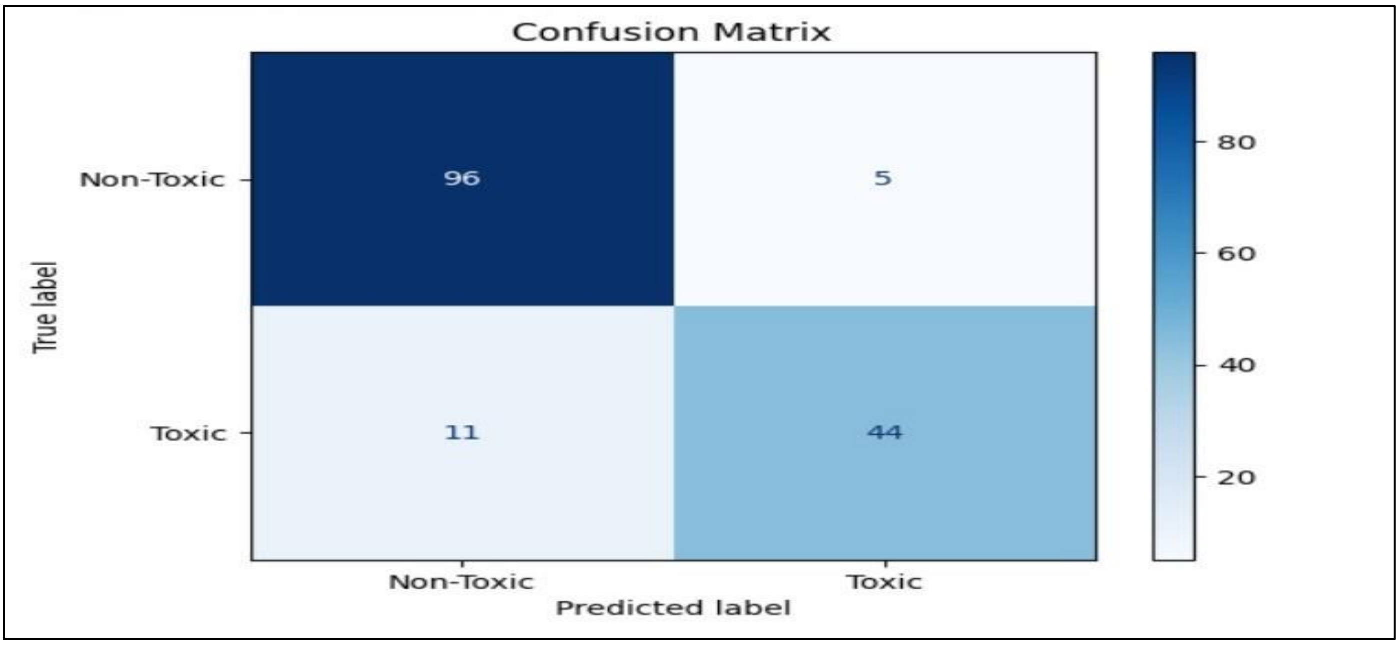

Fig 9 Confusion Matrix for Model with Context

For the model without context, the confusion matrix displays that the model classified 10,440 instances as non-toxic and 10,185 instances as toxic. These results indicate that the model has a high degree of accuracy, as displayed in Figure 6.

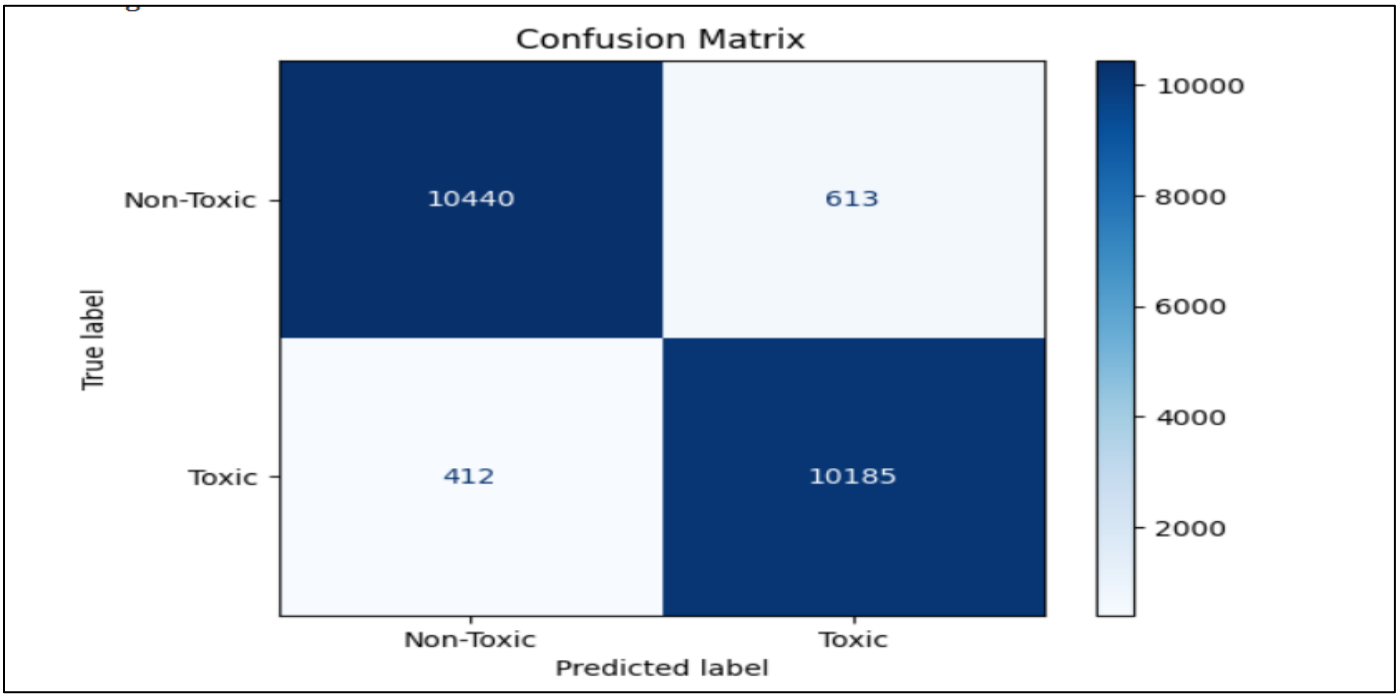

Fig 10 Confusion Matrix for Model without Context

V. CONCLUSION

Our approach addresses important issues that include class imbalance and the need for contextual information and an efficient framework is built which successfully identifies the toxic comments. This approach correctly detects subtle forms of toxicity such as sarcasm and intent, by using two datasets , one with isolated comments and the other with contextual parent comments. Data augmentation techniques using GPT-Neo, along with a fine-tuned RoBERTa model and a weighted ensemble approach makes our system more reliable.

There is a lot of room for improvement, even though our system is quite accurate in detecting the toxicity. This approach would be useful to diverse online communities if it could accommodate datasets concerning different languages. Furthermore, including multimodal data like images and videos could improve the detection of toxicity on different social media platforms. For the purpose of promoting instant moderation and safer, more inclusive online spaces, real-time

ISSN No:-2456-2165 https://doi.org/10.5281/zenodo.14613807

system deployment has to be made possible in the upcoming days. By building on these enhancements, our framework can develop into a proper strategy for maintaining safer and more inclusive online spaces.

REFERENCES

[1]. Chowanda, A., Sutoyo, R., & Tanachutiwat, S. (2021). Exploring text-based emotions recognition machine learning techniques on social media conversation. *Procedia Computer Science, 179*, 821-828.

[2]. Ho, V. A., Nguyen, D. H. C., Nguyen, D. H., Pham, L. T. V., Nguyen, D. V., Nguyen, K. V., & Nguyen, N. L. T. (2020). Emotion recognition for Vietnamese social media text. In *Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11–13, 2019, Revised Selected Papers 16* (pp. 319-333). Springer Singapore.

[3]. Batbaatar, E., Li, M., & Ryu, K. H. (2019). Semantic-emotion neural network for emotion recognition from text. *IEEE Access, 7,* 111866-111878.

[4]. Gaind, B., Syal, V., & Padgalwar, S. (2019). Emotion detection and analysis on social media. *arXiv preprint arXiv:1901.08458*.

[5]. Canales, L., & Martínez-Barco, P. (2014). Emotion detection from text: A survey. In *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC)* (pp. 37-43).

[6]. Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Vora, D., & Pappas, I. (2022). A review on text-based emotion detection—Techniques, applications, datasets, and future directions. *arXiv preprint arXiv:2205.03235*.

[7]. Acheampong, F. A., Wenyu, C., & Nunoo-Mensah, H. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports, 2*(7), e12189.

[8]. Hicham, N., Karim, S., & Habbat, N. (2023). Customer sentiment analysis for Arabic social media using a novel ensemble machine learning approach. *International Journal of Electrical and Computer Engineering, 13*(4), 4504.

[9]. Omuya, E. O., Okeyo, G., & Kimwele, M. (2023). Sentiment analysis on social media tweets using dimensionality reduction and natural language processing. *Engineering Reports, 5*(3), e12579.

[10]. Lian, Z., Liu, B., & Tao, J. (2021). CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 985-1000.

[11]. Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*(7), 5731-5780.

[12]. Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics, 9*(3), 483.

[13]. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), 1093-1113.

[14]. Yue, L., Chen, W., Li, X., Zuo, W., & Yin, M. (2019). A survey of sentiment analysis in social media. *Knowledge and Information Systems, 60*, 617-663.

[15]. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. *Image and Vision Computing, 65*, 3-14.

[16]. Neethu, M. S., & Rajasree, R. (2013). Sentiment analysis in Twitter using machine learning techniques. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.

[17]. Le, B., & Nguyen, H. (2015). Twitter sentiment analysis using machine learning techniques. In *Advanced Computational Methods for Knowledge Engineering: Proceedings of 3rd International Conference on Computer Science, Applied Mathematics and Applications-ICCSAMA 2015* (pp. 279-289). Springer International Publishing.

[18]. Jemai, F., Hayouni, M., & Baccar, S. (2021). Sentiment analysis using machine learning algorithms. In *2021 International Wireless Communications and Mobile Computing (IWCMC)* (pp. 775-779). IEEE.

[19]. Hemalatha, I., Varma, G. S., & Govardhan, A. (2013). Sentiment analysis tool using machine learning algorithms. *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 2*(2), 105-109.

[20]. Mitra, A., & Mohanty, S. (2020). Sentiment analysis using machine learning approaches. In *Emerging Technologies in Data Mining and Information Security* (pp. 63-68). Springer.

[21]. Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P., & Tech, B. (2017). Study of Twitter sentiment analysis using machine learning algorithms on Python. *International Journal of Computer Applications, 165*(9), 29-34.

[22]. Jain, A. P., & Dandannavar, P. (2016). Application of machine learning techniques to sentiment analysis. In *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*(pp. 628-632). IEEE.

[23]. Van Atteveldt, W., Van der Velden, M. A., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures, 15*(2), 121-140.

[24]. Singh, J., Singh, G., & Singh, R. (2017). Optimization of sentiment analysis using machine learning classifiers. *Human-centric Computing and Information Sciences, 7*, 1-12.

[25]. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070.*