Employing TF-IDF and Word2Vec Embeddings to Identify Multi-Class Toxicity through Machine and Deep Learning Approaches

Sanjana Singamsetty
Department of CSE,
SRM University-AP,India
sanjana_s@srmap.edu.in

Harshitha Somayajula
Department of CSE,
SRM University-AP,India
harshitha_s@srmap.edu.in

Anu Likitha Immadisetty
Department of CSE,
SRM University-AP,India
anulikitha_i@srmap.edu.in

Keerthi Sree Konkimalla
Department of CSE,
SRM University-AP,India
keerthisree_k@srmap.edu.in

Srilatha Tokala
ACT Lab, Department of CSE,
SRM University-AP,India
srilatha_tokala@srmap.edu.in

Murali Krishna Enduri
ACT Lab, Department of CSE,
SRM University-AP,India
muralikrishna.e@srmap.edu.in.

Abstract — Our investigation delves into the intricate catalysts triggering toxicity within online discourse, revealing how seemingly innocuous comments can unexpectedly provoke hostile reactions. Given the profound influence of social media viewpoints on individuals, mitigating toxicity emerges as a critical imperative. To address this challenge, we present a sophisticated multi-label classification framework integrating TF-IDF and Word2Vec methodologies for robust vectorization. This framework amalgamates fundamental textual data with intricate metrics derived from prior research, facilitating nuanced monitoring of sentiment shifts, topic dynamics, and conversational context. Leveraging a diverse array of algorithms, including Logistic Regression, AdaBoost, Naive Bayes, Gradient Boosting, as well as Neural Network architectures like LSTM and Bi-LSTM, our model showcases exceptional efficacy in identifying four distinct types of toxicity: 'toxic', 'obscene', 'insult', and 'non-toxic'. Importantly, our study underscores the necessity of accounting for contextual subtleties and sentiment fluctuations in online interactions, advocating for the widespread adoption of advanced natural language processing techniques to foster constructive discourse and enhance digital engagement. Furthermore, our research underscores the dynamic nature of online conversations, emphasizing the need for adaptable frameworks capable of capturing evolving patterns of toxicity.

*Keywords—**Toxicity, Multi-Label Classification, TF-IDF, Word2Vec, Machine Learning, LSTM, Bi-LSTM, Deep Learning***

I. INTRODUCTION

The proliferation of technological sites such as Instagram, Facebook, and YouTube have altered the ways that individuals connect and establish teams around common interests [1,2]. Increased connection, however, provides considerable issues in content control, which is critical for preserving respectful debate. The unprecedented rise in online participation has unavoidably resulted in a spike in toxic remarks, which include abuse, impolite behaviour, and other damaging messages. These unpleasant encounters not only hamper healthy conversation, but they also develop a general culture of distrust, which may have a severe influence on customer satisfaction and even lead to some users entirely disengaging from these platforms. Furthermore, as membership in these virtual networks increases, the work of regulating debates to prevent incivility becomes increasingly difficult. The ongoing problem of toxic material, manifested in many forms of targeted abuse, degrades the quality of online interactions and adds to disputes, consequently reducing the overall user experience[2].

The emergence of online toxicity, which encompasses actions like bullying, and violence present serious challenges for electronic interactions [2,3]. In order to tackle these issues, counter speech and community guidelines have been suggested as well as technological solutions like automated detection and moderation systems to control the large amount of communication that normally occurs in vibrant online communities. Because toxicity on digital platforms may spread quickly and one unpleasant comment can lead to more toxic encounters, research into what causes toxic conduct is crucial. Knowing these triggers is essential to creating interventions that can break these loops at the start and stop online harassment from getting worse [4]. Uncontrolled internet poisoning has serious repercussions that affect people in ways that go far beyond disagreeable interactions. Therefore, it is crucial to identify and address harmful content before it is uploaded, as such materials are making the internet a dangerous space and adversely affecting individuals [5]. Such poisoning can worsen to the point of significant psychological discomfort, which can result in problems like suicidal thoughts, extreme social isolation, harm to one's reputation, and even violent threats.

Academic inquiry into detecting online toxicity has revealed a significant research gap concerning the foundational causes of such toxic interactions [2,4]. It is crucial to comprehend the triggers that initiate toxic dialogues in order to devise effective strategies that mitigate these harmful exchanges and safeguard individuals from their adverse effects. These negative interactions can stifle constructive dialogue, propagate a hostile atmosphere, and lead to a decline in platform engagement, potentially causing users to disengage completely [6]. This requires a targeted investigation into the elements that precipitate toxic conversations, emphasising

the necessity for comprehensive research to accurately identify and counteract these triggers.

To deal with these problems, effective systems that can detect and control harmful information are critical for maintaining the quality of discussions. Recent research has utilised developments in the fields of NLP and machine learning to develop automated systems that can detect harmful comments [7]. However, these attempts frequently lack a detailed understanding of the triggers of toxic responses, which is vital for designing preventative solutions that could avoid the escalation of negative encounters.

Our research provides critical knowledge and methods for creating healthier digital environments, which are critical for sustaining productive and polite interactions across several social media platforms. Our approach to tackle this obstacle entails using powerful natural language processing (NLP) and machine learning techniques to create a strong system capable of recognising harmful information before it becomes apparent to the general public. This proactive strategy aims to prevent the escalation of harmful interactions by identifying and addressing them at their inception. By integrating sophisticated algorithms such as machine learning models and neural networks with deep textual analysis, We want to understand the subtleties and triggers of poisonous comments. This insight is critical for developing effective solutions that can automatically filter out hazardous content, ensuring the integrity and safety of online discussions.

Overall, our method provides a scalable and effective response to the growing concern about online toxicity, enabling a safer and more inclusive digital ecosystem for global consumers. We hope to establish a complete framework that not only detects but also helps avoid the recurrence of such behaviours by a thorough examination of harmful behavioural patterns. This research not only enhances the science of text categorization, but it also has practical implications for enhancing online platform moderation processes, hence improving user experience and community well-being.

II. Related Work

This study builds on previous work targeted at predicting toxicity, as well as the several other toxicity-related markers covered in this material. It reveals a significant increase in demand for study into internet hatred and toxicity, as evidenced by the growing number of peer-reviewed articles on the subject [1]. According to a 2016 poll, toxicity is a major problem, with 66% of reported harassment incidents taking place on social media platforms, prompting 21% of individuals targeted to forgo all forms of social media [2,8]. According to this analysis, toxicity might discourage people from participating in online debates and have a negative impact on their online social connections. Over the last few years, sentiment categorization about toxicity has been a focus of research, particularly in the context of social media data. Researchers have used a variety of machine learning techniques to tackle both the difficulty of toxicity and the more established goal of sentiment analysis. The investigation of comment abuse categorization began with Yin et al.'s pioneering work, which combined TF-IDF with sentiment and contextual information. Their study compared this sophisticated model to a simple TF-IDF technique. When evaluated on chat-style datasets from sites such as Kongregate and MySpace, the findings revealed a significant boost in the classifier's performance. Other research focusing on cyberbullying reflect similar effects [2]. Emphasising how toxicity greatly reduces people's capacity to participate in online debates. Data is crucial for detection, because the data provided determines which labels are used. The researchers develop a machine learning-based system that can identify obscene language in online user comments.

Describes a way for detecting statements of hatred and internet content by focusing on abusive language targeted at group characteristics like ethnicity or religion [8]. It describes the gathering and annotation of a hate speech corpus, as well as techniques for detecting evasion attempts. Salminen et al. (2018) examined 137,098 comments on social media and online news platforms, identifying politicians and media people as top hate speech targets [9]. They used LSTM with GloVe embeddings and obtained an accuracy of 82.5%. Similarly, Spiros et al. (2018) demonstrated encouraging results with CNN on the same dataset [10,11]. This research employs machine learning to categorize tweets into Clean, Offensive, or Hateful groups, integrating various data sources for classification. The methodology encompasses feature extraction and training processes, achieving an accuracy of 87.4% in binary classification (for identifying offensive tweets)and a 78.4% accuracy rate in ternary classification for discerning between hateful, offensive, or clean tweets. Gitari et al. compiled terms from popular hate speech websites and classified them as strongly hateful, slightly hateful, or non-hateful [12]. They used semantic and grammatical patterns to classify a test set, resulting in an F1-score of 65.12.

Significantly, recent studies have highlighted the effectiveness of integrating deep learning methods with word embeddings to achieve superior outcomes in identifying harmful content [13]. These advanced algorithms eliminate the necessity for manual feature selection by autonomously identifying intricate patterns within the data provided. In the realm of natural language processing, word embeddings have gained prominence as a key approach for encoding text, particularly in tasks like Sentiment Analysis, where they consistently exhibit their effectiveness. Moreover, this combination of deep learning and word embeddings not only enhances the accuracy of identifying harmful content but also emphasizes the versatility of these techniques across various language processing tasks.

While past research has mostly focused on detecting and classifying toxic comments, our work takes a novel approach by identifying the drivers of toxic online debates. We provide a sophisticated understanding of the various forms of toxicity found in online conversations by categorising remarks as Toxic, Obscene, Insult, or Non-Toxic. Furthermore, we analyse and classify text data employing various computational methods like TF-IDF and Word2Vec embeddings [14]. Our findings not only address an open and important research subject, but they also have practical ramifications across multiple domains, offering light on the underlying causes of toxic behaviour in online discourse.

III. DATASETS INFORMATION

The dataset for this research was obtained from the Jigsaw Toxic Comment Classification Challenge on Kaggle [15], which included comments from Wikipedia talk page modifications. Human raters manually annotated these remarks to identify toxic behavior into four categories: 'toxic', 'obscene', 'insult', and 'non-toxic'. Each category is denoted by Boolean values (0 or 1) that indicate the presence or absence of the appropriate toxicity kind. To balance the dataset for categorization, we added a new 'non-toxic' label based on the lack of toxic behaviors. The dataset contains over 1 lakh comments, with the following label distribution: 15,294 were classified as 'toxic', 1,595 as 'severe toxic', 8,449 as 'obscene', 478 as 'threat', 7,877 as 'insult', 1,405 as 'identity hate', and 143,346 as 'non-toxic'. Furthermore, 5,707 comments are labelled as 'toxic undefined' and 931 as 'soft toxic', providing additional granularity. 80% of the data (about 1.20 lakh instances) was used for model training, with the remaining 20% (roughly 30,000 instances) reserved for testing.

A. ***Preprocessing***

During pre-processing, the textual data underwent critical procedures to enhance its quality for analysis. Initially, text was tokenized to separate words, then converted to lowercase for uniformity. Lemmatization reduced words to basic forms, standardizing and reducing dimensionality. Stopwords were removed to prioritize meaningful content, and refined further from lemmatized tokens. Additionally, statistics like total tokens, sentences, and punctuation marks provided insight. Overall, these methods ensured clean, standardized data, primed for analysis. The generated data frame was used to generate the TF-IDF matrix and Word2Vec embeddings separately.

1) **TF-IDF**

The TF-IDF (Term Frequency-Inverse Document Frequency) approach is widely used in natural language processing to translate text data into numerical representations suitable for machine learning. It determines the relevance of each word by comparing its frequency within a document to its frequency across all documents in the dataset. This method aids in the identification of terms that are distinctive and significant to individual texts, while decreasing the value of common words such as 'the' or 'and'. In this study, TF-IDF vectors are useful in extracting relevant features for predicting toxicity labels like 'toxic', 'non-toxic', 'obscene', and 'insult'. This technique characterises the linguistic content of each document, allowing for more accurate modelling of toxicity prediction tasks.

2) **Word2Vec**

Word2Vec is a frequently used natural language processing algorithm that converts words into dense vectors in a continuous vector space. This technique makes it easier to represent words with comparable meanings in vector space. In our study, we use the Word2Vec model to identify semantic word associations in textual data. By training Word2Vec on a huge text corpus, it learns to assign numerical vectors to words based on their context. These embeddings provide semantic information that is useful for analyzing word associations, discovering similarities, and assisting with activities like categorizing text and analyzing sentiment. In this work, Word2Vec embeddings improve textual data comprehension and the performance of models used to predict toxicity labels such as 'toxic', 'non-toxic', 'obscene', and 'insult'. This approach enables the model to capture subtle semantic meanings inherent in the text, leading to more precise toxicity predictions.

IV. METHODOLOGY

Our approach entails categorising content with four separate classifications. To do this, we used a robust architecture that combines two critical embedding techniques: TF-IDF and Word2Vec. These embeddings offer the basis for capturing semantic and contextual information in text data. Furthermore, we used a variety of machine learning and deep learning approaches, including logistic regression [16], Naive Bayes [17], Adaboost [18], LSTM [19], Bi-LSTM [20], and Gradient Boosting models. This comprehensive approach allows us to accurately estimate the toxicity of text samples from a variety of categories.

3.A. DEFINITIONS

1. Logistic Regression :

In our comprehensive examination of text-based data, we used logistic regression, a robust and transparent method designed for binary classification problems [15]. We extended this method to its multinomial counterpart in order to successfully manage the complexities of multi-class categorization. This complex model computes the likelihood of each class based on the input features and delivers detailed insights by interpreting the coefficients. We used advanced word embedding techniques, especially Word2Vec and TF-IDF, to improve the model's performance even more. This integration dramatically improved categorization capabilities, resulting in exact text segmentation into four unique groups. Impressively, the combination of logistic regression and Word2Vec embeddings obtained a 95% accuracy, while the implementation with TF-IDF embeddings achieved a significant 87% accuracy.

2. Naive Bayes

In our extensive analysis, we organized textual data into four distinct groups using the Naive Bayes method, a fundamental and extremely effective probabilistic classification approach [19]. We began by thoroughly preparing the text to ensure that it was ready for analytical processing. We subsequently utilized advanced algorithms to convert text into numerical data, notably TF-IDF and Word2Vec, which are critical for preparing textual information for Naive Bayes classifiers. The technique uses Bayes' theorem to estimate the conditional probability for each class based on their intrinsic qualities. Surprisingly, the application of Naive Bayes with TF-IDF yielded an accuracy of 85%, which was commensurate with the findings obtained using Word2Vec embedding. This uniformity in performance across many different kinds of feature representation demonstrates the Naive Bayes algorithm's robustness and reliability in managing numerous data conversion strategies.

3. Ada Boost

In our extensive research, we used AdaBoost, a powerful ensemble learning method that improves prediction accuracy by successively training weak classifiers on data that is adaptively weighted depending on prior classification mistakes [17]. Over repeated rounds, the algorithm improves its accuracy by gradually raising the weights of erroneously identified instances. In this approach, decision stumps served as the foundational weak learners, and their collective outputs

were combined using weighted averaging to produce a decisive categorization. To guarantee that the data was properly structured for AdaBoost, substantial preprocessing was required, particularly to turn textual data into numerical representations such as TF-IDF and Word2Vec embeddings. This phase was necessary to meet the algorithm's requirements for precise and effective classification tasks. Our experimental findings were impressive: AdaBoost achieved 84% accuracy with TF-IDF embeddings and a staggering 95% accuracy with Word2Vec embeddings. These findings illustrate the algorithm's effectiveness and adaptability to several feature representation approaches.

4. Gradient Boosting

Gradient Boosting [21], a powerful ensemble method, was employed to classify textual data into four categories by leveraging both Word2Vec and TF-IDF embeddings. Originally designed for regression, Gradient Boosting has been adapted for classification tasks by iteratively creating decision trees to minimize loss functions, with Word2Vec achieving an impressive 95% accuracy, surpassing the 84% accuracy obtained with TF-IDF. Despite its strengths in capturing complex data interactions, Gradient Boosting struggles with sequential and contextual nuances, which are critical in NLP tasks, making it less effective compared to models like LSTMs or Transformers. Additionally, its high computational cost and tendency to overfit noisy or unstructured text data present challenges, especially when dealing with high-dimensional, sparse features such as those produced by TF-IDF.

5. Long Short-Term Memory (LSTM)

The LSTM model achieves 93.5% accuracy in predicting multiple toxicity labels ('toxic', 'non-toxic', 'obscene', 'insult') using a two-layer architecture with 64 and 32 units and dropout regularization [21]. Input data, represented as TF-IDF vectors, undergoes sequence reshaping for LSTM processing, optimized with binary cross-entropy loss and Adam optimizer over 10 epochs. Word2Vec embeddings, trained similarly, achieve 96.5% accuracy by capturing semantic relationships and language nuances, positioning similar words close in a multidimensional space. This enhanced context understanding contributes to superior toxicity prediction compared to TF-IDF. TF-IDF, emphasizing word importance and corpus frequency, offers contextual depth but may exhibit slightly lower accuracy in complex tasks.

6. Bi-Long Short Term Memory (Bi-LSTM)

The BI-LSTM model employs two bidirectional LSTM layers (64 and 32 units) with dropout (0.3) to predict four toxicity labels ('toxic', 'obscene', 'insult', 'non-toxic') simultaneously. Input data is converted into TF-IDF vectors and Word2Vec embeddings to capture word relationships and semantic meanings. During training, binary cross-entropy loss and Adam optimizer are utilised over 10 epochs in a batch size of 32.2, alongside validation data for monitoring and preventing overfitting. Sigmoid activation facilitates multi-label classification. Word2Vec training employs binary cross-entropy and Adam optimizer over 15 epochs with the same batch size. Word2Vec achieved 84% average accuracy, while TF-IDF achieved 93.5% on the four labels, demonstrating robust performance in toxicity prediction .Both the BiLSTM (Bidirectional LSTM) and LSTM models perform similarly when trained on TF-IDF vectors, yielding comparable accuracy on toxicity label prediction tasks.

V. PROPOSED MODEL

The framework shown in Figure 1 depicts a thorough approach to categorising text into four distinct categories: poisonous, obscene, insult, and nontoxic. This method combines established techniques like TF-IDF (Term Frequency-Inverse Document Frequency) with cutting-edge methodology like Word2Vec embeddings, machine learning, and deep learning algorithms, as shown in Fig. 1.

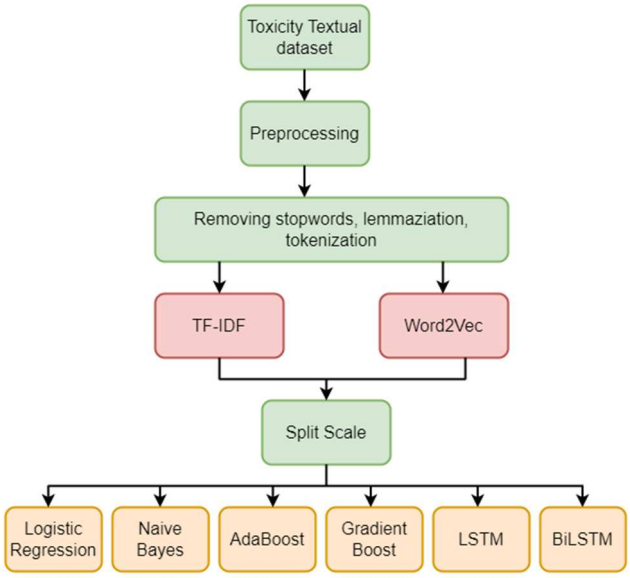

Fig.1. A snippet of the proposed approach procedure

Initially, a diversified dataset with classifications such as toxic, obscene, insult and non-toxic was gathered. The acquired data was then preprocessed, including stop word removal and punctuation cleanup, to improve its quality and relevancy. Following preprocessing, two different strategies for feature extraction were used: TF-IDF transformation and Word2Vec embeddings. The TF-IDF transformed data was then classified using a variety of machine learning techniques such as AdaBoost, Logistic Regression, Gradient Boosting, and Naive Bayes. Similarly, the Word2Vec embeddings were combined with the same machine learning methods for categorization. In addition, deep learning methods, notably Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), were used with both TF-IDF converted data and Word2Vec embeddings to capture detailed patterns and dependencies in text. Throughout the process, feature scaling was used to ensure consistency and optimal performance across all algorithms. Finally, the trained models were tested in terms of accuracy, precision, recall, and F1-score to determine their usefulness in categorising text based on toxicity and content.

VI. RESULTS

In our research study, we used a variety of evaluation measures to for classifying toxicity in social media. To gauge the effectiveness of deep learning and machine learning models, we analysed a variety of performance criteria, each of which provided unique insights into the models' ability to reliably forecast toxicity in comments. These measurements include the following [21]:

Test Accuracy: To measure a model's overall performance, compute the percentage of cases correctly predicted out of the total number of instances as shown in Eq.1.

$$Accuracy(a) = \frac{K+L}{K+L+M+N} \quad (1)$$

Precision: Precision relates to how well a model predicts positive outcomes. It is calculated by reducing the number of true positives by the sum of true positives and false positives, resulting in the percentage of successfully detected positives among all positive predictions as shown in Eq.2.

$$Precision(p) = \frac{K}{K+M} \quad (2)$$

Recall: Recall is a measure that evaluates the frequency with which a machine learning model correctly detects positive instances (true positives) out of all the genuine positive samples in the dataset as shown in Eq.3.

$$recall(r) = \frac{K}{K+N} \quad (3)$$

F1-Score: It's a single score that shows indicates the overall effectiveness of the model. It combines precision and recall into one number, giving an idea of how well the model deals with mistakes. as shown in Eq.4.

$$F1 - Score(r) = \frac{2(p*r)}{p+r} \quad (4)$$

In the equations, K, L, M, and N represent different outcomes of classification. K is the number of true positives (correct positive predictions), L is the number of true negatives (correct negative predictions), M is the number of false positives (incorrect positive predictions), and N is the number of false negatives (incorrect negative predictions).

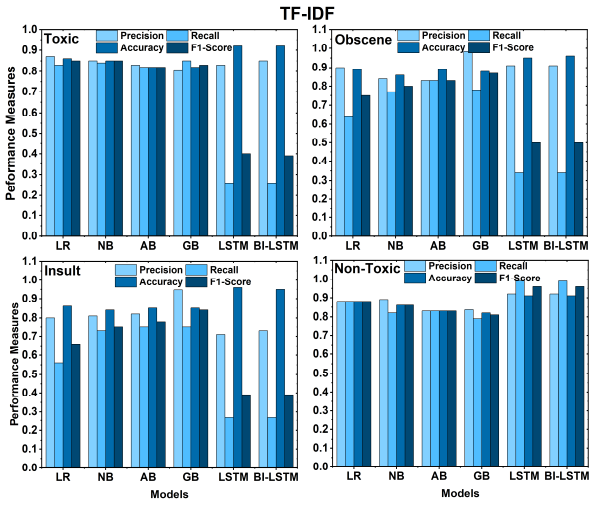

Fig.2. Assessment criteria for different techniques used in the toxicity dataset by using TF-IDF

Figure 2 shows the Assessment criteria for different techniques used in the 'toxicity.csv' TF-IDF vectorization was used to create the dataset. The chart shows that for all assessment measures, the LSTM algorithm and the BI-LSTM method both have a 93% success rate, performs well compared to the other algorithms mentioning the reason due to giving weight to the uniqueness of words in relation to the data processing capability of LSTM and BI-LSTM based models due to their design for sequence data. The least performance is given by the Gradient Boosting, though a potent ML Method that can model complex patterns in data, it doesn't perform on text data vectorized by TF-TDF because it does not naturally consider the sequential and contextual aspects of language. Gradient Boosting treats features independently, which can be a limitation in text classification tasks where the sequence of words is important. While, other algorithms may not be designed to handle the sequential nature of text data or the nuanced importance IF-IDF assigns to each word.

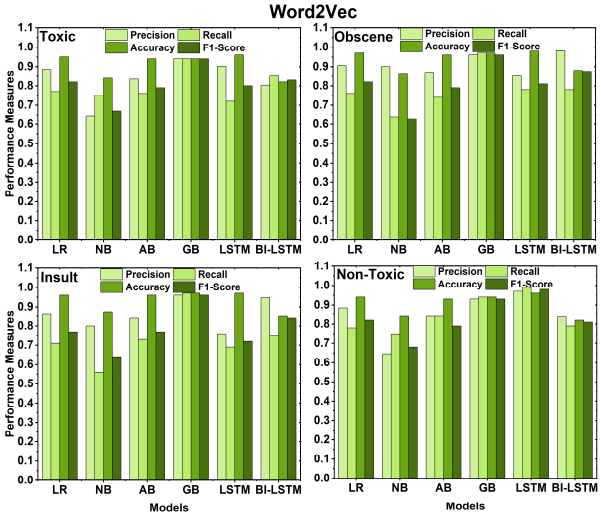

Fig.3 Assessment criteria for different techniques used in the toxicity dataset by using Word2Vec

Figure 3 illustrates the effectiveness of hybrid methodologies in our analysis, utilizing Word2Vec vectorization. It demonstrates that the LSTM model beats other algorithms with an accuracy of 92%, but the BI-LSTM model exhibits the lowest performance at 84%, identifying the explanation as Word2Vec, which provides dense word embeddings that capture semantic meanings by capitalizing on the surrounding word context, LSTM can exploit the sequential context provided by Word2Vec embeddings to understand the pattern or sequence in the text which explains the high accuracy rate observed as it is effectively uses the contextual information encoded in Word2Vec to make accurate predictions and Additionally, training BI-LSTMs can be more challenging; they have more parameters to learn, which might not have been optimized as well as the LSTM in the particular case.

VII. Conclusion and Future Work

In the current study, toxicity was categorized into four distinct labels: Toxic, Obscene, Insult, and Non-Toxic. These classifications underwent meticulous evaluation using two word embedding techniques, along with assessments of LSTM and BiLSTM models employing the same embeddings. Additionally, ensemble methods like AdaBoost and Gradient Boosting were employed. The comprehensive analysis provided valuable insights into effective text classification methods. Notably, the study highlighted the effectiveness of ensemble methods and logistic regression, emphasizing their role in achieving high classification accuracy. Importantly, Word2Vec embeddings coupled with LSTM achieved the highest accuracy. Logistic Regression also performed exceptionally well, surpassing ensemble methods. These findings underscore the importance of meticulous model selection and feature engineering, particularly when integrating word embedding techniques like Word2Vec. Furthermore, the study's extensive use of

multi-label classification with a large dataset yielded promising results, especially when combined with the presented approaches. Overall, the study sheds light on the optimal strategies for accurate text classification, emphasizing the effectiveness of ensemble methods and logistic regression, with Word2Vec embeddings coupled with LSTM emerging as the top performer. The aforementioned results provide useful recommendations for future research aiming at improving text categorization performance in a variety of applications.

In addition to the study's thorough analysis, one exciting a potential direction for future research is to explore the integration of deep learning models with attention mechanisms for toxicity classification. Attention approaches have proven effective across numerous natural language processing challenges by enabling models to focus on relevant regions of the input sequence. Adding attention mechanisms to LSTM or Bi-LSTM architectures may improve the model's ability to recognize slight nuances in hazardous language, resulting in higher classification accuracy. Furthermore, investigating the application of transfer learning methodologies, such as adjusting pre-trained language models such as BERT or GPT could help to progress the field by using large-scale language representations learnt from massive text corpora. Furthermore, given the changing nature of language and internet communication, Creating adaptive models that can continuously learn from incoming data streams and adjust to changing language patterns would be beneficial. This could include strategies like online learning or incremental training, which ensure that toxicity classification models are current and effective in real-world circumstances. Future research endeavours that investigate these directions can build on the study's findings and contribute to the continued enhancement of text classification performance across a variety of applications.

VIII. References

[1]. Salminen, J., Almerekhi, H., Milenković, M., Jung, S. G., An, J., Kwak, H., & Jansen, B. (2018, June). Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 12, No. 1).

[2]. Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2020, April). Are these comments triggering? predicting triggers of toxicity in online discussions. In *Proceedings of the web conference 2020* (pp. 3033-3040).

[3]. Almerekhi, H., Kwak, H., Jansen, B. J., & Salminen, J. (2019, September). Detecting toxicity triggers in online discussions. In *Proceedings of the 30th ACM conference on hypertext and social media* (pp. 291-292). Del Bosque, L. P., & Garza, S. E. (2016). Prediction of aggressive comments in social media: an exploratory study. *IEEE Latin America Transactions*, *14*(7), 3474-3480.

[4]. Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12-18.

[5]. Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. New Media & Society, 18, 410 - 428.

[6]. Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. CSCW : proceedings of the Conference on Computer-Supported Cooperative Work. Conference on Computer-Supported Cooperative Work, 2017, 1217–1230. https://doi.org/10.1145/2998181.2998213.

[7]. Yin, D., Xue, Z., Hong, L., Davison, B.D., & Edwards, L. (2009). Detection of Harassment on Web 2.0.

[8]. Stevens, F., Nurse, J. R., & Arief, B. (2021). Cyber stalking, cyber harassment, and adult mental health: A systematic review. Cyberpsychology, Behavior, and Social Networking, 24(6), 367-376

[9]. Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., & Plagianakos, V. P. (2018, July). Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th hellenic conference on artificial intelligence* (pp. 1-6).

[10]. Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, *6*, 13825-13835.

[11]. Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215-230.

[12]. Dessì, D., Recupero, D. R., & Sack, H. (2021). An assessment of deep learning models and word embeddings for toxicity detection within online textual comments. *Electronics*, *10*(7), 779.

[13]. Hosseinmardi, H., Mattson, S. A., Ibn Rafiq, R., Han, R., Lv, Q., & Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the Instagram social network. In Social Informatics: 7th International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings 7 (pp. 49-66). Springer International Publishing.

[14]. Aseervatham, S., Gaussier, É., Antoniadis, A., Burlet, M., & Denneulin, Y. (2012). Logistic regression and text classification. Textual Information Access: Statistical Models, Part-II.

[15]. Kaggle. *Jigsaw toxic comment classification challenge*. Kaggle. Retrieved October 21, 2024, from https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/code.

[16]. Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

[17]. Xu, S. (2018). Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*, *44*(1), 48-59.

[18]. Chengsheng, T. U., Bing, X. U., & Huacheng, L. I. U. (2018, May). The application of the AdaBoost algorithm in the text classification. In *2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)* (pp. 1792-1796). IEEE.

[19]. Hassan, S. U., Ahamed, J., & Ahmad, K. (2022). Analytics of machine learning-based algorithms for text classification. *Sustainable operations and computers*, *3*, 238-248.

[20]. Zhou, H. (2022). Research of text classification based on TF-IDF and CNN-LSTM. In *journal of physics: conference series* (Vol. 2171, No. 1, p. 012021). IOP Publishing.

[21]. Rahul, Kajla, H., Hooda, J., & Saini, G. (2020). Classification of Online Toxic Comments Using Machine Learning Algorithms. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), 1119-1123.