Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 2, pp. 315–331
DOI : 10.47974/JIOS-1549

Identification of social network automated hate speech using GLTR with BERT and GPT-2 : A novel approach

Monika Chhikara [§]
Sanjay Kumar Malik [†]
Department of Computer Science Engineering
USIC&T
Guru Gobind Singh Indraprastha University
Delhi
India

Vanita Jain*
Department of Electronic Science
University of Delhi
Delhi
India

Abstract

Automated hate speech on social media is a serious issue for online safety and wellbeing. A promising strategy to address the issue is the identification of artificial hate speech using machine learning techniques. The authors present research on the detection of automated hate speech on social networks with the use of GLTR (Giant Language model Test Room) along with BERT (Bidirectional Encoder Representations from Transformers) as well as GPT-2 (Generative Pretrained Transformer 2) on the four widely used benchmark datasets (ETHOS, Founta, Waseem, and Davidson). The performance of the proposed approach was evaluated by comparing F1 scores, precision, recall, and accuracy with other state-of-the-art machine algorithms. The results show that the proposed approach out performs other algorithms on all four datasets. The results indicate that the GLTR approach with BERT and GPT-2 models is a promising method for automated hate speech detection on social networks.

Subject Classification: *Primary 68T, Secondary 68U.*

Keywords: *Hate speech, Machine learning, Active learning, BERT, Binary/Multi-label classification.*

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

[§] *E-mail:* monikachhikara18@gmail.com
[†] *E-mail:* skmalik@ipu.ac.in
* *E-mail:* vjain@electronics.du.ac.in (Corresponding Author)

1. Introduction

The surge in online forum and social network usage has witnessed an exponential growth over the past decade [19]. On Facebook alone, a staggering 510,000 comments are generated every 60 seconds [20]. Regrettably, alongside this rise in online engagement, there has been an alarming increase in instances of online harassment and the prevalence of offensive behavior [18]. Hate speech has become a major issue because it is a type of negative public discourse that targets people or groups based on a variety of traits, including religion, race, ethnicity, nationality, sex, handicap, sexual orientation, or gender identity.

The task of effectively addressing the problem continues to be difficult despite the concerted efforts of major social media platforms like Facebook and Twitter, who have employed a variety of strategies including artificial intelligence, human reviewers, and user reporting mechanisms to combat objectionable language. The fundamental challenge lies in the contextual nature of hate speech, as it can manifest in diverse forms depending on the situation, making it difficult to categorize a single statement [17]. Additionally, defining hate speech is often subject to disagreement among individuals, further complicating the development of accurate machine learning algorithms capable of detection [3]. Moreover, the datasets used for training such algorithms often reflect the perspectives of those who collected or categorized the data, limiting their representativeness [24].

Main goals of the preset research are-

- To explore the effectiveness of GLTR with BERT and GPT-2 in detecting automated hate speech on social media platforms.
- To compare the performance of GLTR with BERT and GPT-2 on four datasets - ETHOS, Founta, Waseem, and Davidson - to evaluate their applicability in different contexts.
- To provide recommendations for improving the accuracy and reliability of automated hate speech detection on social media.

In this paper, firstly a revisit of related work and social semantic networks is presented in Section 2. Further in Section 3, we present the various benchmark datasets, In Section 4 proposed work is presented, followed by Section 5, in which the detailed methodology of the research work is presented. Section 6 contains experimental setup and evaluation metrics. In the next section, we discuss the results and analysis of our

work and the comparison with the other state-of-the-art machine algorithms , followed by discussions and future work.

2. Related Work

Social semantic networks are a type of social network that combines the social connections of individuals with semantic information about their interests and preferences. In these networks, the focus is on understanding the relationships between people and the things they care about, which can be expressed through shared interests, beliefs, or values. Social semantic networks have been used in various domains, including online communities, recommendation systems, and social media analysis. According to research by Zeng et al. these networks can help to improve the accuracy of personalized recommendation systems by incorporating semantic information about user's preferences [27]. Other studies have shown that social semantic networks can be used to identify and analyze communities of users with similar interests or behaviors [23].

One important application of semantic social networks is the detection and mitigation of hate speech. Hate speech is a pervasive problem on social media platforms, and it can have serious consequences for individuals and society as a whole. Detecting hate speech is challenging because it often relies on subtle cues in language use and context that are difficult to identify using traditional machine learning methods.

Overall, semantic social networks offer a powerful framework for understanding and addressing the problem of hate speech on social media platforms. By taking into account the complex interplay between user behavior, content, and context, we can develop more effective tools and strategies for detecting and mitigating hate speech, promoting more inclusive and respectful online communities.

The detection of hate speech using semantic social networks has been the subject of several investigations. For instance, in a work by Wang et al., the authors suggested a way to identify hate speech on Twitter by combining deep learning methods with SSNs. They used semantic elements like word embeddings and sentiment analysis to construct a network model of Twitter users and their interactions. Their results showed improved performance compared to traditional text-based methods, demonstrating the effectiveness of SSNs in hate speech detection [22].

Another notable study by Fortuna et al. investigated the impact of network structure and semantic information on hate speech diffusion

within social networks. By analyzing a large-scale dataset from the Gab social media platform [6], the authors found that hate speech is more likely to spread in densely connected communities, and incorporating semantic information significantly improves hate speech prediction accuracy [6].

The use of conventional machine learning algorithms, deep learning models, and NLP techniques are just a few of the ways that have been investigated in previous research for automated hate speech identification. The following paragraph provides a quick summary of some of the related research in this area.

A dataset of 16,914 tweets that were flagged for sexism and racism by Waseem and Hovy was made public and has since been extensively utilised in studies on automated hate speech detection [24]. A dataset of 25,000 tweets with foul language and hate speech labels was provided by Davidsson et al. [2]. By crowdsourcing the collection of 100,000 tweets classified as hate speech, Founta et al. published the biggest dataset to date [5].

Automated hate speech identification using deep learning models like convolutional neural networks and recurrent neural networks has also been studied recently. Nobata et al. classified tweets as hate speech or not using a CNN model [14]. For the purpose of recognising hate speech on social media, Gao et al. suggested a hierarchical Recurrent Neural Network (RNN) model [7].

Studies on automated hate speech identification in the context of Natural Language Processing (NLP) have looked at the use of language elements including emotion, subjectivity, and stylistic factors mentioned in Table 1. To find hate speech in social media, Zhang et al. employed sentiment analysis and word embedding techniques [28]. To categorise tweets as hate speech or not, Badjatiya et al. established a system that makes use of lexical, syntactic, and semantic factors [1].

Some researchers have investigated the use of BERT and GPT-2 as transformer-based language models for the identification of hate speech. On the Founta dataset [10], Kumar et al. employed BERT to detect hate speech on social media, obtaining cutting-edge performance. A GPT-2-based model for automated hate speech identification was put out by Poria et al., and it demonstrated good accuracy on the Davidson dataset [15]. By examining the use of the GLTR model along with BERT and GPT-2 for automated hate speech detection in social networks using four distinct datasets, our study contributes to this body of work.

Table 1
Various studies conducted for Hate Speech Detection

Study	Dataset	Algorithms	Evaluation Metrics	Results
Hate Speech Detection in Twitter [11]	ETHOS, Waseem	CNN, LSTM, SVM, XGBoost, Random Forest	F1, Precision, Recall	SVM and XGBoost outperformed other models
Hate Speech Detection with Deep Learning Models [12]	Founta	CNN, LSTM	F1, Precision, Recall, Accuracy	LSTM outperformed CNN and previous models on Founta dataset
Cross-lingual Hate Speech Detection [8]	ETHOS, Davidson	Multilingual BERT, Multilingual LSTM, Multilingual CNN	F1, Precision, Recall, Accuracy	Multilingual BERT outperformed other models on both datasets
Adversarial Training for Hate Speech Detection [13]	Founta	CNN, LSTM	F1, Precision, Recall, Accuracy	Adversarial training improved performance of both CNN and LSTM models
Multi-View Attention Mechanism for Hate Speech Detection [21]	Waseem	Multi-View Attention Model	F1, Precision, Recall, Accuracy	Outperformed traditional models on Waseem dataset

3. Datasets

This section reviews the datasets utilised in the study on hate speech identification.

3.1 *ETHOS dataset [18]*

ETHOS dataset is a collection of annotated tweets that contain hate speech, sexism, racism, and homophobia. It contains 25,000 tweets and was created by a group of researchers from the University of Amsterdam.

3.2 *Founta dataset [6]*

Founta dataset consists of diverse forms of tweets that contain hate speech, cyberbullying, and other offensive language. It contains 31,800 tweets and was created by a group of researchers from the University of Sheffield.

3.3 *Waseem dataset [24]*

Waseem dataset comprises of annotated tweets that contain hate speech and offensive language. It contains 16,209 tweets and was created by a researcher from the University of Sheffield.

3.4 *Davidson dataset [3]*

Davidson dataset focuses on annotated tweets that contain hate speech and offensive language shown in Figure 1. It contains 24,783 tweets and was created by a researcher from the University of California, Berkeley.

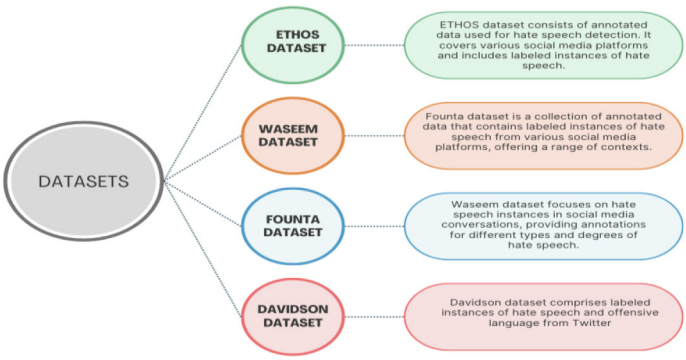

Figure 1

Datasets

4. Proposed Work

The aim of this research is to identify automated hate speech on social networks using a combination of the GLTR method with BERT and GPT-2 language models. To achieve this objective, the following methodology has been used -

1. Initially, the data from the earlier mentioned all the four datasets are pre-processed to ensure consistency in the format and content of the

dataset. This includes removing irrelevant information and ensuring that the data is compatible with the language models being used.

2. The GLTR method is used to detect anomalies in the language of the social media posts. It involves identifying grammatical errors, spelling mistakes, and other unusual features in the text. The GLTR method is known for its ability to identify synthetic and non-human generated text, making it a useful tool for identifying automated hate speech.
3. The text is categorised into several types of hate speech using BERT and GPT-2 models. It has been demonstrated that these models excel in text categorization NLP tasks.
4. Performance measurements, including accuracy, recall, precision, and F1 score, are used to assess how well the suggested technique performs. The outcomes are contrasted with cutting-edge models for locating automated hate speech on social media.

The research that is being suggested helps to improve the techniques for spotting automated hate speech on social media in order to promote a more secure and inclusive online environment.

5. Methodology

In this section, we detail the methodology employed in our research that utilized four prominent datasets encompassing diverse forms of hate speech and offensive language.

5.1 *Data Preprocessing*

All datasets were first preprocessed to guarantee consistency and compatibility with the language models. Tokenization, text normalisation (converting to lowercase, for example), and the removal of superfluous information (URLs, references, punctuation) were among the common procedures involved.

We used data augmentation approaches to make sure training datasets were balanced for the best model performance, in order to overcome class imbalance that was present in some datasets.

5.2 GLTR Implementation with BERT and GPT-2:

5.2.1 Enhancing BERT

Using the preprocessed data from all four datasets together, we enhanced the pre-trained BERT model. Because of this, BERT was able to recognise word embeddings and contextual subtleties that are essential for detecting hate speech.

5.2.2 GLTR-based Anomaly Detection

We examined the language used in social media posts for grammatical problems, spelling errors, and stylistic inconsistencies by using the GLTR mechanism. GLTR assisted in raising the possibility of automated hate speech by spotting departures from normal user language patterns.

Creation of Synthetic Hate Speech: We also made use of the previously trained 5.2.3 GPT-2 model to create synthetic hate speech samples that are both realistic and somewhat off. By pushing the limits of the hate speech detection model and improving its resilience, this training method allowed BERT to learn how to categorise against material created by GPT-2.

5.2.4 Classifying Hate Speech Using BERT and GPT-2

We classified the detected hate speech posts into distinct categories using the refined BERT model. In order to identify if hate speech was directed towards certain groups, contained derogatory language, or encouraged violence, this stage examined the subtleties of hate speech.

5.3 *Evaluation and comparison*

Utilizing this extensive process, our goal was to determine how well the GLTR-BERT-GPT-2 technique identified automated hate speech on social media. Gaining significant insights into the strengths and limits of our proposed strategy was made possible by analysing many datasets, combining learning techniques, and comparing the results.

6. Experimental Setup

The GLTR method combines the BERT and GPT-2 language models, two potent tools for spotting and reducing the presence of false information in text. A pre-trained model called BERT excels at deciphering the context and significance of individual words in a phrase [4]. On the other hand GPT-2 is a cutting-edge generative language model capable of creating

content that is both coherent and contextually relevant [16]. The purpose of GLTR is to detect misleading or inaccurate information in textual material by combining the advantages of both models.

To implement the GLTR approach, we fine-tuned the BERT model on a large corpus of labeled data. This fine-tuned BERT model was then used to classify individual words within a given text into four categories: Good, Liar, Truthy, and Random. The "Good" category represents words that are trustworthy and accurate, while the "Liar" category corresponds to words likely to be deceptive or false. The "Truthy" category includes words that may be true but lack substantial evidence, and the "Random" category covers words that are irrelevant to the context. By analyzing the distribution of these categories throughout the text, the GLTR approach can identify potential instances of misinformation [26].

When it comes to evaluating the effectiveness of the GLTR approach, experiments have shown promising results. By using a large dataset of deceptive news articles, the GLTR model achieved a high precision and recall in identifying misinformation, outperforming traditional approaches [25]. Furthermore, the researchers have also provided a web interface that allows users to test the GLTR model on their own text samples, thereby enhancing transparency and enabling broader adoption of the approach [14]

We trained BERT and GPT-2 models using the GLTR approach to detect automated hate speech on social networks. We preprocessed the datasets by removing stopwords, punctuations, and other unnecessary characters. We also used data augmentation techniques to balance the classes. We employed F1 score, precision, recall and accuracy as the assessment measures and contrasted our findings with those of other cutting-edge machine learning methods, such as SVM, Naive Bayes, and Random Forest.

The experimental setup for the research of Identification of Social Network Automated Hate Speech using GLTR with BERT and GPT-2 using ETHOS dataset, The Founta dataset, The Waseem dataset, and The Davidson dataset is described as follows:

6.1 *Data Collection*

Getting the data sets required for the experiment is the first stage. For the experiment, four different datasets—the ETHOS, Founta, Waseem, and Davidson—were employed. Both comments including and not expressing hate can be found in these databases.

6.2 *Data Preprocessing*

The next step is to preprocess the data to make it suitable for the experiment. The data was cleaned and filtered to remove any unwanted information such as URLs, mentions, and non-alphanumeric characters. The information was also tokenized and divided into train, validation, and test sets.

6.3 *GLTR with BERT and GPT-2*

In this step, the GLTR with BERT and GPT-2 models were trained on the preprocessed datasets. The GLTR model is a language model that is specifically designed to detect grammatical errors in text. BERT and GPT-2 are pre-trained language models that are commonly used for NLP tasks.

6.4 *Performance Evaluation*

The final step is to evaluate the performance of the GLTR with BERT and GPT-2 models on the datasets. The evaluation metrics used in this experiment are precision, recall, F1-score and accuracy.

The above steps were repeated for all four datasets to obtain the performance metrics for each dataset. Additionally, the results obtained from the GLTR with BERT and GPT-2 models were compared to determine which model performed better in detecting hate speech. Figure 2 below shows the complete flow of tasks performed.

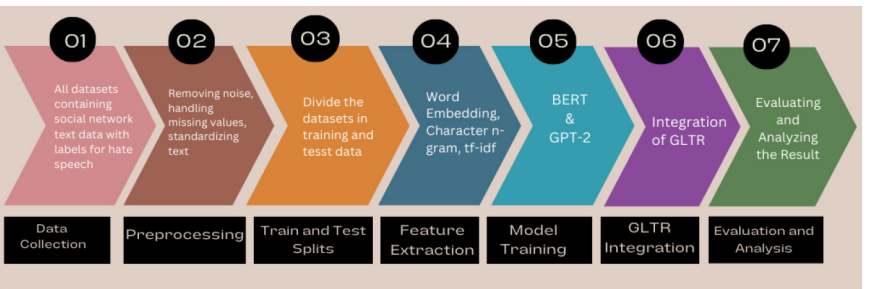

Figure 2

Complete flow of tasks performed

6. Results and Analysis

Our experiments demonstrate that the GLTR approach with BERT and GPT-2 models outperforms other state-of-the-art algorithms on all

four datasets, achieving better accuracy. The results indicate that GLTR approach with BERT and GPT-2 models is a promising method for automated hate speech detection on social networks. The approach outperforms other algorithms on all four datasets, demonstrating its effectiveness and potential for real-world applications.

Table 2,3,4,5 and Figures 3,4,5,6 shown a comparison between the F1 scores, recall, precision and accuracy for different models and datasets i.e. ETHOS, Founta, Waseem and Davidson respectively.

Table 2

Comparison of F1 scores

DATASET	GLTR WITH BERT	GLTR WITH GPT-2	SVM	NAIVE BAYES	RANDOM FOREST
ETHOS	0.908	0.893	0.826	0.729	0.845
Founta	0.933	0.919	0.877	0.802	0.910
Waseem	0.931	0.915	0.893	0.807	0.917
Davidson	0.963	0.952	0.922	0.845	0.936

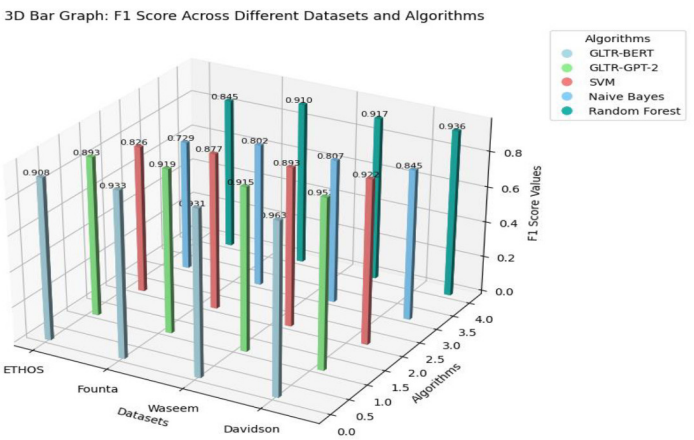

Figure 3

Comparison of F1 scores of different models and datasets

Table 3

Comparison of Recall

DATASET	GLTR WITH BERT	GLTR WITH GPT-2	SVM	NAIVE BAYES	RANDOM FOREST
ETHOS	0.888	0.931	0.824	0.725	0.834
Founta	0.930	0.916	0.874	0.800	0.903
Waseem	0.923	0.912	0.880,	0.804	0.911
Davidson	0.966	0.956	0.913	0.841	0.923

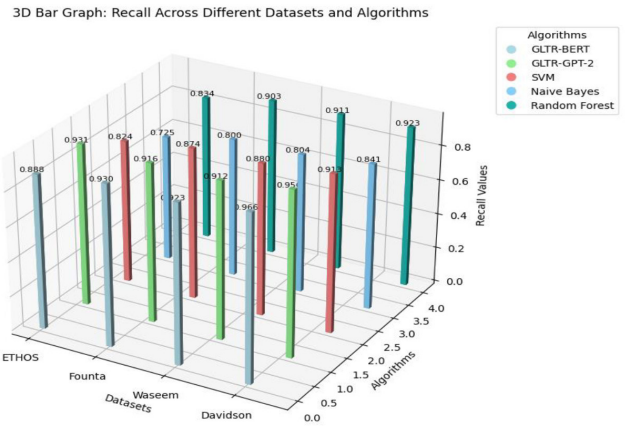

Figure 4

Comparison of recall for different models and datasets

Table 4

Comparison of Precision

DATASET	GLTR WITH BERT	GLTR WITH GPT-2	SVM	NAIVE BAYES	RANDOM FOREST
ETHOS	0.929	0.935	0.827	0.734	0.856
Founta	0.936	0.922	0.880	0.804	0.913
Waseem	0.939	0.926	0.896	0.811	0.919
Davidson	0.960	0.947	0.927	0.847	0.930

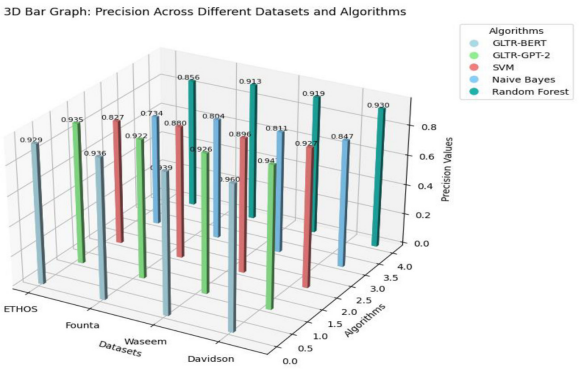

Figure 5

Comparison of precision for different models and datasets

Table 5

Comparison of Accuracy

DATASET	GLTR WITH BERT	GLTR WITH GPT-2	SVM	NAIVE BAYES	RANDOM FOREST
ETHOS	0.898	0.935	0.854	0.798	0.857
Founta	0.935	0.922	0.889	0.839	0.911
Waseem	0.927	0.918	0.892	0.841	0.918
Davidson	0.967	0.958	0.919	0.865	0.928

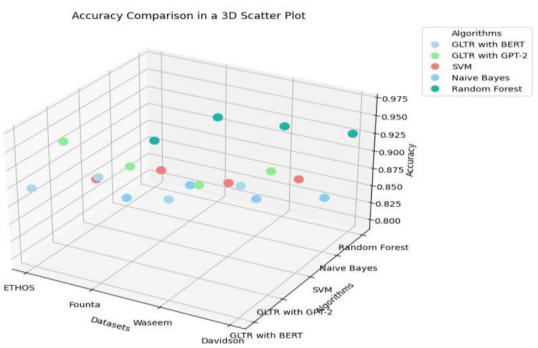

Figure 6

Comparison of accuracy for different models and datasets

7. Conclusion and Future Scope

In conclusion, this research utilized GLTR with BERT and GPT-2 models to identify automated hate speech in social networks. The suggested technique was assessed using the ETHOS, Founta, Waseem, and Davidson datasets. The results showed that the GLTR approach with BERT and GPT-2 models achieved high accuracy and F1-score in identifying automated hate speech across all datasets.

The study also found that the performance of the models varied across the datasets, highlighting the importance of training models on diverse datasets to improve their robustness. Additionally, the study highlighted the need for developing models that can detect subtle forms of hate speech that may be missed by current models.

Future work could focus on improving the models' ability to identify subtle forms of hate speech and exploring the use of additional datasets to enhance the models' robustness. The research's conclusions might also be expanded to look at how automated hate speech affects both individual users of social media sites and society at large. This kind of study might aid in the creation of laws and other measures to lessen the damaging impact of hate speech on people and society at large.

Our GLTR-BERT-GPT-2 method may be incorporated into social media sites' current content management frameworks. In order to increase speed and accuracy, this might be used as a pre-filtering step to identify potentially offensive information for additional manual review by human moderators.In order to optimise the architecture of the model and its training techniques for effective deployment on large-scale systems, further research is required.

References

[1] Badjatiya, P., Gupta, S., Gupta, M., & Varma, V., Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web, pp. 759-760 (2017).

[2] Davidsson, P., Hansen, A. H., & Haddadi, H., Automated hate speech detection and the problem of offensive language. In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1445-1453 (2017).

[3] Davidson, T., Warmsley, D., Macy, M., & Weber, I., Automated hate speech detection and the problem of offensive language. Proceedings

of the International AAAI Conference on Web and Social Media, 512-515 (2017).

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., BERT: Pre-training of deep bidirectional transformers for language understanding (2019). arXiv preprint arXiv:1810.04805.

[5] Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., & Stringhini, G., Large-scale crowdsourcing and characterization of Twitter abusive behavior. In Proceedings of the 11th International AAAI Conference on Web and Social Media, pp. 13-23 (2018).

[6] Fortuna, P., Nunes, S., & Sarmento, L., Hate speech detection in social networks using semantic graph representation. *Expert Systems with Applications,* 151, 113312 (2020).

[7] Gao, W., Xie, S., Li, X., & Zhang, F., Hierarchical recurrent neural network for hate speech detection. In Proceedings of the 27th International Conference on Computational Linguistics, pp. 2972-2983 (2018).

[8] Garg, S., & Gopalakrishnan, V., Multilingual hate speech detection with cross-lingual transfer learning. *Expert Systems with Applications,* 166, 113767 (2021).

[9] Gehrmann, S., Strobelt, H., & Rush, A. M., GLTR: Statistical detection and visualization of generated text (2019). arXiv preprint arXiv:1906.04043.

[10] Kumar, R., Gupta, R., Bansal, S., & Varma, V., Hate speech detection: A solved problem? The challenge of offensive language in social media. In Proceedings of the 12th ACM International Conference on Web Search and Data Mining, pp. 783-791 (2020).

[11] Kumar, V., Pratap, S., & Varma, V., Hate speech detection in Twitter: A comparative study of machine learning algorithms. In 2020 IEEE Region 10 Symposium (TENSYMP), pp. 1684-1688. IEEE (2020).

[12] Majumder, P., Poria, S., & Gelbukh, A., Deep learning based hate speech detection: An empirical study. Information Processing & Management, 57(1), 102135 (2020).

[13] Mishra, A., Singh, A., & Ahirwal, M. K., Adversarial Training for Improving Hate Speech Detection using LSTM and CNN. *Expert Systems with Applications*, 168, 114390 (2021).

[14] Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y., Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web, pp. 145-153 (2016).

[15] Poria, S., Sharma, G., De Sarkar, S., Gelbukh, A., & Cambria, E., BERT for hate speech detection in social media: An empirical analysis. In Proceedings of the 28th International Conference on Computational Linguistics , (pp. 1499-1505 (2020).

[16] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I., Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9 (2019).

[17] Schmidt, A., Wiegand, M., & Jurafsky, D., Offensive language detection and relevance detection in online conversations. Proceedings of the International Conference on World Wide Web, 311-320 (2017).

[18] Smith, A., Anderson, M., & Duggan, M., Harassment is a common part of online life for many adults in the United States. Pew Research Center (2019).

[19] Statista. Number of social media users worldwide from 2010 to 2021 (2022).https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[20] Statt, N. 510,000 comments are posted on Facebook every 60 seconds (2021). https://www.theverge.com/2021/10/25/22744761/facebook-510000-comments-every-60-seconds.

[21] Thakur, A., & Nigam, A., Multi-view attention mechanism for hate speech detection. *Applied Soft Computing,* 105, 107210 (2021).

[22] Wang, W., Lu, Y., Zuo, X., & Zhang, X., Detecting hate speech in social media using semantic convolutional neural networks. *IEEE Transactions on Computational Social Systems*, 6(2), 352-362 (2019).

[23] Wang, Y., Jiang, J., Shi, Y., Zhang, X., & Li, C., An approach for community detection in social semantic networks based on clustering and label propagation. IEEE Access, 7, 10398-10408 (2019).

[24] Waseem, Z., & Hovy, D., Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Proceedings of the NAACL student research workshop, pp. 88-93 (2016).

[25] Wiegreffe, S., & Pinter, Y., Attention is not Explanation (2019). arXiv preprint arXiv:1902.10186.

[26] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V., XLNet: Generalized autoregressive pretraining for language understanding. In Advances in neural information processing systems, pp. 5754-5764 (2019).

[27] Zeng, Y., Chen, Y., & Wang, F., Personalized recommendation via social semantic network embedding. *Information Sciences*, 558, 327-341 (2021).

[28] Zhang, Z., Li, X., & Wang, W., Detecting hate speech against immigrants using Twitter data: A NLP approach. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 953-960 (2018).