Comparative Analysis of Hate Speech Detection: Traditional vs. Deep Learning Approaches

Haibo PEN[2], Nicole TEO[1], Zhaoxia WANG[1*]
School of Computing and Information Systems, Singapore Management University, Singapore[1]
Key Laboratory of Smart Grid of Ministry of Education, Tianjin University, Tianjin 300072, China[2]
penhaibo@tju.edu.cn; nicolet.2023@engd.smu.edu.sg; zxwang@smu.edu.sg*

***Abstract*—Detecting hate speech on social media poses a significant challenge, especially in distinguishing it from offensive language, as learning-based models often struggle due to nuanced differences between them, which leads to frequent misclassifications of hate speech instances, with most research focusing on refining hate speech detection methods. Thus, this paper seeks to know if traditional learning-based methods should still be used, considering the perceived advantages of deep learning in this domain. This is done by investigating advancements in hate speech detection. It involves the utilization of deep learning-based models for detailed hate speech detection tasks and compares the results with those obtained from traditional learning-based baseline models through multidimensional aspect analysis. By considering various aspects to gain a comprehensive understanding, we can discern the strengths and weaknesses in current state-of-the-art techniques. Our research findings reveal the performance of traditional learning-based hate speech detection outperforms that of deep learning-based methods. While acknowledging the potential demonstrated by deep learning methodologies, this study emphasizes the significance of traditional machine learning approaches in effectively addressing hate speech detection tasks. It advocates for a balanced perspective, highlighting that dismissing the capabilities of traditional methods in favor of emerging deep learning-based techniques may not consistently yield the most effective results.**

***Index Terms*—Deep learning, Hate speech detection, Performance comparison, Traditional learning-based methods, Multidimensional aspect analysis**

I. Introduction

Hate speech on social media, targeting individuals based on race, ethnicity, religion, gender, or sexual orientation, has become a topic of significant concern. According to the Anti-Defamation League (ADL), instances of online hate speech have notably escalated, with a reported 52% rise in adult harassment last year compared to 40% in 2022. On the other hand, offensive language is a broader category that includes all forms of insults and vulgarity. As a result, hate speech might be categorized as a particular kind of offensive language. Therefore, to address the issue of hate speech and mitigate societal concerns, there is a pressing need for research into automated hate speech detection methods.

Unfortunately, detecting hate speech is challenging problem due to the lack of a clear and precise definition. Nonetheless, advancements in big data and natural language processing (NLP) have led to the development of various detection techniques [1]–[6]. Such big data and NLP techniques form the basis for various hate speech detection methods. These encompass traditional machine learning techniques [7]–[9] and deep learning approaches [10], [11]. In traditional learning-based methods, Support Vector Machines (SVM), XGBoost, and Naive Bayes (NB) are commonly utilized, typically requiring vector representations of text data [12].

In addition, Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) are commonly used in hate speech detection. Besides, deep learning methods, such as Convolutional Neural Networks combined with Word2Vec, FastText, or Glove embeddings, provide more expressive representations and hierarchical feature learning [13]–[15]. These deep learning approaches have demonstrated high performance in hate speech detection tasks [16].

As discussed above, both traditional and deep learning methods have been employed for hate speech detection tasks. Researchers may find interest in comparing their performance. While deep learning-based methods have demonstrated considerable success in improving hate speech detection tasks, their definitive superiority over traditional methods in certain real-life tweet scenarios remains a topic of debate. This research aims to determine whether traditional machine learning methods should still be used, given the rise in popularity of deep learning approaches in automatic hate speech detection. In this regard, we make the following contributions:

1) The research conducts a comprehensive comparison between traditional learning-based hate speech detection methodologies and deep learning-based methods through multidimensional aspect analysis to identify situations where traditional methods excel in specific real-life hate speech detection tasks.
2) The research offers valuable insights into hate speech detection methodologies. It highlights the challenges in distinguishing hate speech from offensive language and underscores the frequent misclassifications by learning-based models. By delving into advancements and empirical comparisons, the paper provides a deeper understanding of the strengths and weaknesses of different methods in tackling hate speech detection tasks.
3) The research advocates a balanced perspective in methodology choice for hate speech detection, emphasizing the importance of traditional machine learning

*Corresponding Author: Zhaoxia WANG (e-mail: zxwang@smu.edu.sg)

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00070

approaches while acknowledging the potential of deep learning methodologies.

II. Literature Review

Hate speech detection, also known as "automatic hate speech detection", is a subset of sentiment analysis, which involves classifying digital text based on its emotional tone [1], [4], [5], [17]–[19]. Like sentiment analysis, hate speech detection addresses a classification problem [7]–[9], [20], where a classifier assigns a class to input text. Two primary classification approaches are found in the literature: traditional learning-based classifiers and deep learning-based classifiers.

A. Traditional Hate Speech Detection Methods

Most early research in hate speech detection predominantly relied on traditional classifiers, with SVM being the most popular, followed by Logistic Regression (LR), Naive Bayes (NB), and Random Forest (RF), among others [20]. These classifiers are important in the development of hate speech detection and can also lead to more accurate detection of offensive texts. For instance, Davidson et al. [21] disclosed that LR performs better with the appropriate n-gram range of 1 to 3 for the L2 normalization of TF-IDF which has 95.6% accuracy. Their research produced the CrowdFlower dataset, and employed unigrams, bigrams, and trigrams weighted by TF-IDF in their work. In their work, they tried to classify the Twitter data into the categories "hateful", "offensive" and "neither" using the aforementioned features as well as count indicators for hashtags, mentions, and retweets. They fed their features to a variety of classifiers like Linear SVM, LR, NB, RF, and Decision Tree (DT). The results showed the best performing models were Linear SVM and LR [21].

B. Deep learning-based Hate Speech Detection Methods

Deep learning based methods, which have demonstrated success in the hate speech detection domain for their potential in terms of performance. The most popular models are Convolutional Neural Networks (CNN), Long-Short Term Memory Networks (LSTM) and Bidirectional Encoder Representations from Transformers (BERT), among others [22]. For instance, Kim et al. [23], uses CNN with a single convolutional layer and word2vec for word embeddings to classify various short sentences from different databases. The results are quite significant for such a simple model, with the best CNN achieving an F1 score of 0.89. In addition, Badjatiya et al. [24] used CNN, LSTM and gradient boosted decision trees (GBDT) with Glove and random word embeddings to classify text into "sexist", "racist" and "neither". They compared the performance of the deep learning methods to several traditional classifiers with Bag-of-Word vectors, character n-grams, and TF-IDF. The results of each deep learning model showed an F1 score above 0.804. The best performing model was with LSTM, random embeddings, and GBDT which achieved a recall, precision, and F1 score of 0.93.

III. Methodology for Traditional vs. Deep Learning-based Hate Speech Detection

This section outlines the proposed methodology designed for hate speech detection. Specifically, two comparative approaches are employed in this study to achieve optimal performance and compare the results with existing solutions. The block diagram depicted in Figure 1 illustrates the main steps, providing a visual representation of our proposed methodology and the process leading to our final results.

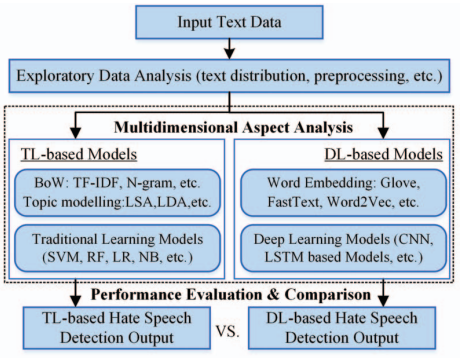

Fig. 1. Architecture overview of the proposed method.

A. Hate Speech Dataset

We used Davidson [21], a widely used dataset, in our experiments. It was created for research on hate speech detection and is used to study and develop algorithms for identifying hate speech, offensive language, and non-offensive content in tweets. The dataset is unique in the sense that the data distinguishes between hateful and offensive language, of which the latter is derogatory but not necessarily a threat to society. As such, it enables investigation into the distinction between hateful and offensive language. It is important to note that this dataset contains text that may be considered racist, sexist, homophobic, or generally offensive. The dataset containing 25,297 records was collected from Twitter and annotated by CrowdFlower users who judged the content of each tweet to determine whether it was "hateful", "offensive" or "neither". However, from the label class distribution statistics indicator, it has been found that there was a massive class imbalance in the dataset with a high proportion of the tweets in the offensive language class at 77.4% while the remaining two classes were underrepresented. In fact, only 5.8% of the tweets are considered hateful, and 16.8% are considered as neither hateful nor offensive. In the evaluation of the real-world context, this makes sense. Although hate speech is increasingly prevalent online, it is much less common than offensive speech.

B. Exploratory Data Analysis

After importing the dataset, exploratory data analysis was performed to better understand the data at hand and acquire important information necessary for determining the next processing steps. This includes text data preprocessing and handling of imbalanced data.

(1) Text data preprocessing

To prepare the data for modeling, we used several natural language processing techniques. Firstly, we cleaned the text data. This entailed the removal of unnecessary data such as usernames, retweets, duplicates, and special characters. In addition, all emojis were converted from Unicode to text. Secondly, we analyzed the text word frequency. It was found that the distinction between hate speech and offensive language is that the former has more social context association i.e. words that are associated with racial or gender stereotypes. Offensive language tweets, on the other hand, contained more curse words. Thirdly, we duplicated the text data. It was found that there were 869 duplicates. We chose to drop the duplicates and only keep the first values. From this, we can see that most duplicates occurred for the hateful and offensive language class as their distribution decreased by 0.1% compared to neither class. Finally, we split the text data. After processing, the data was ready to be split into a train, validation and test set. The respective ratios used were 60%, 20% and 20%.

(2) Handling of imbalanced data

There still remains a big class imbalance with the offensive language class taking up 77.3%, which is an insignificant difference as compared to unprocessed data. Since an imbalanced dataset is hard for a model to train, the problem is addressed in the training dataset using random oversampling. This is a technique in which random points from the minority class are selected and duplicated to increase the number of data points in the minority class. In general, oversampling ensures an equal distribution of data points for all 3 classes in the training dataset. For this paper, random oversampling was more appropriate than undersampling. Undersampling involves the removal of data points in the majority class. This increases the chance of information loss, which is critical to avoid if we are already working with a smaller dataset.

C. Multidimensional aspect analysis

(1) Traditional learning-based models

For traditional learning-based methods, feature extraction is essential. CountVectorizer (CV) and TF-IDF Vectorizer are among the most popular methods. CountVectorizer transforms text into vectors based on word frequency, while TF-IDF Vectorizer calculates the relevance of a word in a given text. To comprehensively assess the capabilities of traditional machine learning methods, a multidimensional aspect analysis was conducted, comparing both CV and TF-IDF Vectorizer.

To assess the potential of traditional learning methods in hate speech detection, we established a benchmark model for comparison. Serving as baseline models, several traditional machine learning algorithms, including LinearSVC (a variant of SVM) [25], LR [26], RF [27], and NB [28], were selected based on their prominence in the literature. These models were trained using processed data to evaluate their effectiveness. Both CV and TF-IDF Vectorizer were used to extract features for training these baseline models. This multidimensional aspect analysis is used to determine the most suitable feature extraction method. By employing various traditional classifiers and feature extraction techniques, this study aimed to establish a comparative benchmark against which the performance of models in hate speech detection could be evaluated.

(2) Deep learning-based models

In contrast, deep learning-based methods utilize word embeddings instead of traditional feature extraction processes. These embeddings encapsulate the semantic meaning of words, representing data points in a less sparse, lower-dimensional space, thereby enhancing model training. Glove and Word2Vec are widely-used pre-trained word embeddings, developed by Stanford and Google, respectively, trained on extensive datasets. To perform multidimensional aspect analysis, both Glove and Word2Vec, specifically using the skip-gram method for Word2Vec, were employed and compared. Additionally, an embedding based on the corpus generated by all tweets was trained for performance comparison.

Due to the excellent results achieved with a simple one-layer CNN for short sentence classification, we decided to follow a similar architecture and emphasize the use of CNNs. The applicability of Kim's model [23] was justified by considering the overlap between short sentences and Twitter tweets. For all embeddings, a dimension of 100 was used. Although Kim used an embedding of 300 for his work on short sentence classification, our dataset is smaller in comparison, which thus requires a smaller dimensionality. In addition, bigger dimensionalities can capture more of the semantic meaning and can lead to overfitting, which is a significant factor to consider if the dataset is not too big. Overall, two CNN models were developed in this paper: 1) A simple single channel CNN [23], [29]. The model in this paper entails that a single convolutional layer was used with one kernel size; 2) A multi-channel CNN [29], [30]. The model in this paper is a multichannel CNN as it uses three concatenated convolutional layers, each with a differently sized kernel. As such, it enhances the possibility to discover more different features from the text embedding layer, since each kernel considers a different number of words at a time as it slides over text data.

The multidimensional aspect analysis thoroughly examines both traditional learning and deep learning methods, providing a comprehensive understanding of their strengths and weaknesses in addressing hate speech detection tasks.

IV. Experiment Results and Discussion

This section includes a detailed description of the used dataset and hate speech detection technical details of each step in traditional vs. deep learning approaches.

A. Evaluation Metrics

Performance evaluation of hate speech detection models typically makes use of the classic precision P_r, recall R_c and F1-score F metrics [7]. Although precision is commonly used, it is not applicable for this paper due to the imbalance in the dataset. Therefore, recall and F1 score were used to evaluate the models. The F1 score provides a better metric to assess overall classification ability. In addition, recall was

included because it measures all the positives and how many the system predicts as positive. As the distinction between hate and offensive language is not well defined, it is important to build a classifier that can clearly predict hate language versus offensive language. As such, we aim to minimize the number of false negatives. For both F1 score and recall, the weighted and macro averages were included. This is because the class imbalance can make the weighted average less informative, whereas the macro average can inform clearer on how the model performs in classifying the minority classes. Additionally, because the focus of this paper is on hate speech, we looked at the recall of hate speech and to maximize that.

B. Multidimensional Performance Comparative Analysis

(1) Traditional machine learning models

We investigated the performance of SVM (linearSVC), LR, RF, and NB as a classifier. All traditional machine learning models have different feature extraction methods - CV and TF-IDF. These models were developed and fine-tuned through Gridsearch/Randomsearch to optimize their performance in the classification task. The results of the comparative analysis are shown in Table I and Figure 2.

TABLE I
EVALUATION METRICS VALUE WITH TRADITIONAL MODELS

Model	Evaluation Metrics					
	Recall value			F1 value		
	Macro	Weighted	(Hate)	Macro	Weighted	(Hate)
SVM+CV	0.70	0.85	0.42	0.68	0.86	0.33
SVM+CV+Gridsearch	0.65	0.83	0.33	0.64	0.84	0.26
SVM+TF-IDF	0.73	0.85	0.47	0.70	0.86	0.37
SVM+TF-IDF+Gridsearch	0.65	0.84	0.31	0.65	0.85	0.26
LR+CV	0.76	0.87	0.48	0.72	0.88	0.38
LR+CV+Randomsearch	0.69	0.84	0.38	0.67	0.85	0.29
LR+TF-IDF	0.78	0.87	0.55	0.73	0.88	0.42
LR+TF-IDF+Randomsearch	0.72	0.87	0.40	0.70	0.87	0.35
RF+CV+Unigram	0.71	0.89	0.32	0.71	0.89	0.37
RF+CV+Bigram	0.55	0.63	0.16	0.47	0.67	0.22
RF+TF-IDF+Unigram	0.70	0.90	0.25	0.71	0.89	0.34
RF+TF-IDF+Bigram	0.55	0.63	0.17	0.47	0.66	0.21
NB+CV+Unigram	0.73	0.84	0.53	0.69	0.86	0.34
NB+CV+Bigram	0.54	0.67	0.50	0.49	0.73	0.18
NB+TF-IDF+Unigram	0.70	0.81	0.56	0.65	0.83	0.32
NB+TF-IDF+Bigram	0.56	0.66	0.53	0.50	0.72	0.19

From Table I and Figure 2, LR with TF-IDF can be seen as the best performing model. Although random forest performs better for the weighted recall and weighted F1 score, the difference is marginal – 0.02 for the weighted recall and 0.01 for the weighted F1 score. Furthermore, for all other metrics, LR outperforms all other models. The most significant result

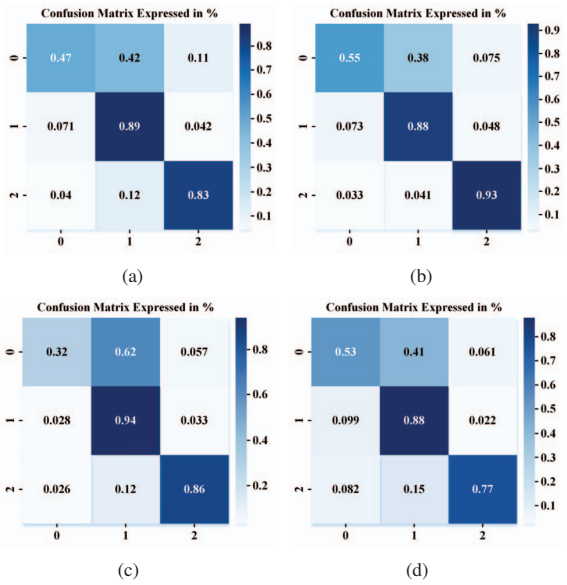

Fig. 2. Confusion matrix of best traditional learning-based models, (a)SVM with TF-IDF vectorizer; (b)LR with TF-IDF vectorizer; (c)RF with CountVectorizer; (d)NB with CountVectorizer.

is that the F1 score for detecting hatred is 0.42, 0.05 points higher than the second best performing models.

(2) Deep learning-based models

We investigated the performance of single channel CNN, and multi channel CNN as a classifier. All CNN models with different embedding methods - self-trained, Word2Vec and Glove. The detail results of the comparative analysis are shown in Table II and 3.

TABLE II
EVALUATION METRICS VALUE WITH DEEP LEARNING MODELS

Model	Evaluation Metrics					
	Recall value			F1 value		
	Macro	Weighted	(Hate)	Macro	Weighted	(Hate)
Single Channel+self-trained	0.65	0.87	0.24	0.66	0.86	0.28
Multi Channel+self-trained	0.70	0.85	0.42	0.68	0.86	0.34
Single Channel+Word2Vec	0.68	0.84	0.41	0.67	0.85	0.33
Multi Channel+Word2Vec	0.71	0.86	0.37	0.68	0.86	0.32
Single Channel+Glove	0.69	0.87	0.35	0.69	0.87	0.35
Multi Channel+Glove	0.75	0.87	0.49	0.72	0.88	0.41

Table II and Figure 3 show the results of the different CNN models on the test set. The first striking observation is the differences between the weighted and macro values for recall and F1. That is, the weighted recall and F1 values are quite high, with a range of 0.84 to 0.87 for recall and 0.85 to 0.88 for F1 score. Considering both values together, they are an indication that the models are quite successful classifiers.

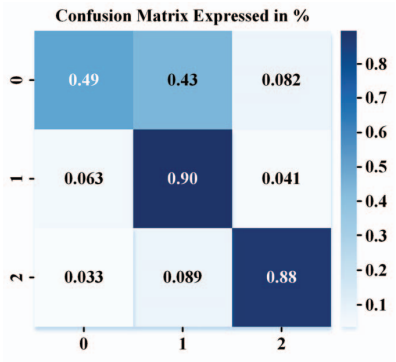

Fig. 3. Confusion matrix of best deep learning-based model.

TABLE III
EVALUATION METRICS VALUE WITH BEST TRADITIONAL VS DEEP LEARNING MODELS

Model	Evaluation Metrics					
	Recall value			F1 value		
	Macro	Weighted	(Hate)	Macro	Weighted	(Hate)
SVM+TF-IDF	0.73	0.85	0.47	0.70	0.86	0.37
LR+TF-IDF	0.78	0.87	0.55	0.73	0.88	0.42
RF+CV+Unigram	0.71	0.89	0.32	0.71	0.89	0.37
NB+CV+Unigram	0.73	0.84	0.53	0.69	0.86	0.34
Multi Channel+Glove	0.75	0.87	0.49	0.72	0.88	0.41

However, when looking at the macro values for recall and F1 score, we can see that they are much lower. For recall values from 0.65 to 0.75 and for F1 scores from 0.66 to 0.72. Macro averages do not consider the size of the classes, so they are a truer indication of how well the model does if one considers each class equivalently important. Again, as for the other classifiers, the culprit for the discrepancy between macro and weighted values for recall and F1 score is the hate class. Specifically, the classifier has a very hard time at accurately classifying hate speech. This is indicated by the recall and F1 score values for the hate class, which are even lower than those for the macro values of recall and F1. The lowest value for recall of the hate class is 0.24 for the single channel CNN with a self-trained embedding. This means that of all the hate speech, the classifier managed to only correctly classify 24%.
(3) Traditional vs Deep learning Comparative Analysis

Table III further compares the best traditional baseline model and the best deep learning model. In absolute terms, logistic regression can be seen as the best model, performing higher or equally as high as the deep learning model across all evaluation metrics. Interestingly, considering all metrics, the deep learning model significantly underperforms the most in the recall for the hate class. While the logistic regression model has a value of 0.55 for the recall of the hate class, the deep learning model has a value of 0.49 for the same metric. As such, the logistic regression model is better at classifying hate speech. Nevertheless, this difference should not be valued too much when considering general performance. Both models still struggle with classifying hate speech as more than 45% and 51% are misclassified as offensive language for the logistic regression model and deep learning model respectively. Therefore, given the results of the problem in the introduction, the logistic regression model is the best performing model.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

The current paper investigated the use of traditional learning-based methods in comparison to deep learning-based methods in the task of hate speech detection. Several traditional machine learning algorithms were tested to create a robust baseline model to compare the deep learning model against through multidimensional aspect analysis. Among the traditional machine learning models: SVM, LR, RF, NB with unigrams and TF-IDF proved to be the best scoring one with recall (macro) 0.78, recall (weighted) 0.87, recall (hate) 0.55, F1 (macro) 0.73, F1(weighted) of 0.88 and F1(hate) of 0.42. For the deep learning models, the emphasis was on CNNs as inspired by Kim's work on short sentence classification. The best performing model was multichannel CNN with Glove word embedding, with a score of recall (macro) 0.75, recall (weighted) 0.87, recall (hate) 0.49, F1 (macro) 0.72, F1(weighted) 0.88, and F1(hate) 0.41.

All models had considerably lower results for the macro metrics compared to the weighted metrics, which was due to the difficulty the models had in classifying hate speech. In particular, most of the models misclassified a significant portion of the hate speech as offensive language. This illustrates the fine line between hate speech and offensive language, and room for improvement on building a classifier that can more clearly distinguish them. Most models did well on classifying offensive language, which can be attributed to it being the majority class.

Comparing the best traditional baseline model and deep learning model, the overall best model was the baseline model, logistic regression. Therefore, in response to the problem statement, traditional machine learning methods should still be used given the rise in popularity of deep learning approaches in automatic hate speech detection. Although the best deep learning and baseline models' performance was close, the traditional baseline model still outperformed the deep learning model. In addition, what must also be taken into consideration is the ease of use of logistic regression. Contrary to the CNN with Glove embedding, significantly fewer parameters had to be tuned. As such, although deep learning methods are on the rise, we must not neglect the power of traditional machine learning methods in automatic hate speech detection.

B. Future Work

For future work, we could include social media posts from different languages and platforms to create a dataset with a

variety of linguistic and cultural contexts. This is due to the possibility of distinct, context-specific hate speech patterns appearing on various platforms and language groupings. The models can be trained to be more culturally sensitive and broadly applicable by expanding the range of data. We could collect and annotate information from different social media sites, forums, and comment sections throughout the globe. To guarantee proper representation and annotation, linguists and cultural experts would need to work together on this.

To enhance hate speech identification efficacy, exploring and optimizing various deep learning architectures is crucial. Models such as CNNs and LSTMs have shown promise in text categorization tasks. However, hybrid models and architectural tuning could further improve performance. Systematic hyperparameter adjustments, including kernel sizes, layer numbers, and types, can effectively leverage both local and sequential text features. Moreover, leveraging recent advancements in Large Language Models (LLMs) presents opportunities for enhancing hate speech detection accuracy, warranting further investigation in future research.

ACKNOWLEDGMENT

The authors express their sincere appreciation to the following SMU students for their enthusiastic interest and invaluable contributions to this Speech Emotion Analysis research: Bodine Stubbé, Sara Knapp, Ong Jun Yang, Bryan Koh Kai Yu, Eunice Ong Yujie, Edlyn Tan Tse Yin, Tan Zuyi Joey, Bodine Salomi Marie-Louise Stubbe, Shambhavi Goenka, Darryl Soh Soon Yong, Chung Zhi Huai, and Elston Eng Shi Yang. Their dedication is indispensable for this paper.

REFERENCES

[1] Z. Wang, C. S. Chong, L. Lan, Y. Yang, S. B. Ho, and J. C. Tong, "Fine-grained sentiment analysis of social media with emotion sensing," in *2016 Future technologies conference (FTC)*. IEEE, 2016, pp. 1361–1364.

[2] A. Gandhi, P. Ahir, K. Adhvaryu, P. Shah, R. Lohiya, E. Cambria, S. Poria, and A. Hussain, "Hate speech detection: A comprehensive review of recent works," *Expert Systems*, p. e13562, 2024.

[3] Z. Wang, Z. Hu, S.-B. Ho, E. Cambria, and A.-H. Tan, "Mimusa–mimicking human language understanding for fine-grained multi-class sentiment analysis," *Neural Computing and Applications*, vol. 35, pp. 15 907–15 921, 2023.

[4] Z. Hu, Z. Wang, Y. Wang, and A.-H. Tan, "MSRL-Net: A multi-level semantic relation-enhanced learning network for aspect-based sentiment analysis," *Expert Systems with Applications*, vol. 217, p. 119492, 2023.

[5] Z. Wang, V. J. C. Tong, and H. C. Chin, "Enhancing machine-learning methods for sentiment classification of web data," in *Information Retrieval Technology: 10th Asia Information Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014. Proceedings 10*. Springer, 2014, pp. 394–405.

[6] Z. Wang, Z. Hu, F. Li, S.-B. Ho, and E. Cambria, "Learning-based stock trending prediction by incorporating technical indicators and social media sentiment," *Cognitive Computation*, vol. 15, pp. 1092–1102, 2023.

[7] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: a review," *IEEE Access*, vol. 9, pp. 88 364–88 376, 2021.

[8] I. Aljarah, M. Habib, N. Hijazi, H. Faris, R. Qaddoura, B. Hammo, M. Abushariah, and M. Alfawareh, "Intelligent detection of hate speech in arabic social network: A machine learning approach," *Journal of Information Science*, vol. 47, no. 4, pp. 483–501, 2021.

[9] F. E. Ayo, O. Folorunso, F. T. Ibharalu, I. A. Osinuga, and A. Abayomi-Alli, "A probabilistic clustering model for hate speech classification in twitter," *Expert Systems with Applications*, vol. 173, p. 114762, 2021.

[10] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in arabic tweets using deep learning," *Multimedia systems*, vol. 28, no. 6, pp. 1963–1974, 2022.

[11] R. Duwairi, A. Hayajneh, and M. Quwaider, "A deep learning framework for automatic detection of hate speech embedded in arabic tweets," *Arabian Journal for Science and Engineering*, vol. 46, pp. 4001–4014, 2021.

[12] S. Abro, S. Shaikh, Z. H. Khand, A. Zafar, S. Khan, and G. Mujtaba, "Automatic hate speech detection using machine learning: A comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 484–491, 2020.

[13] A. Toktarova, D. Syrlybay, B. Myrzakhmetova, G. Anuarbekova, G. Rakhimbayeva, B. Zhylanbaeva, N. Suieuova, and M. Kerimbekov, "Hate speech detection in social networks using machine learning and deep learning methods," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 5, pp. 396–406, 2023.

[14] K. Maity, S. Bhattacharya, S. Saha, and M. Seera, "A deep learning framework for the detection of malay hate speech," *IEEE Access*, vol. 11, pp. 79 542–79 552, 2023.

[15] N. Badri, F. Kboubi, and A. H. Chaibi, "Combining fasttext and glove word embedding for offensive and hate speech text detection," *Procedia Computer Science*, vol. 207, pp. 769–778, 2022.

[16] D. Lin, "Appliance of deep learning on hate speech detection: A systematic review," *Highlights in Science, Engineering and Technology*, vol. 31, pp. 76–81, 2023.

[17] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 3829–3839.

[18] Z. Wang, S.-B. Ho, and E. Cambria, "Multi-level fine-scaled sentiment sensing with ambivalence handling," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 28, no. 4, pp. 683–697, 2020.

[19] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "SenticNet 7: A commonsense-based neurosymbolic AI framework for explainable sentiment analysis," in *LREC*, 2022, pp. 3829–3839.

[20] Z. Mansur, N. Omar, and S. Tiun, "Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities," *IEEE Access*, vol. 11, pp. 16 226–16 249, 2023.

[21] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, no. 1, 2017, pp. 512–515.

[22] H. Saleh, A. Alhothali, and K. Moria, "Detection of hate speech using bert and hate speech word embedding with deep model," *Applied Artificial Intelligence*, vol. 37, no. 1, p. 2166719, 2023.

[23] H. Kim and Y.-S. Jeong, "Sentiment classification using convolutional neural networks," *Applied Sciences*, vol. 9, no. 11, p. 2347, 2019.

[24] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in *Proceedings of the 26th international conference on World Wide Web companion*, 2017, pp. 759–760.

[25] R. Khezzar, A. Moursi, and Z. Al Aghbari, "arhatedetector: detection of hate speech from standard and dialectal arabic tweets," *Discover Internet of Things*, vol. 3, no. 1, pp. 1–13, 2023.

[26] P. S. B. Ginting, B. Irawan, and C. Setianingsih, "Hate speech detection on twitter using multinomial logistic regression classification method," in *2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*. IEEE, 2019, pp. 105–111.

[27] K. Nugroho, E. Noersasongko, A. Z. Fanani, R. S. Basuki *et al.*, "Improving random forest method to detect hatespeech and offensive word," in *2019 International Conference on Information and Communications Technology (ICOIACT)*. IEEE, 2019, pp. 514–518.

[28] K. K. Kiilu, G. Okeyo, R. Rimiru, and K. Ogada, "Using naïve bayes algorithm in detection of hate tweets," *International Journal of Scientific and Research Publications*, vol. 8, no. 3, pp. 99–107, 2018.

[29] Z. Abebaw, A. Rauber, and S. Atnafu, "Design and implementation of a multichannel convolutional neural network for hate speech detection in social networks," *Revue d'Intelligence Artificielle*, vol. 36, no. 2, pp. 175–183, 2022.

[30] M. Fazil, S. Khan, B. M. Albahlal, R. M. Alotaibi, T. Siddiqui, and M. A. Shah, "Attentional multi-channel convolution with bidirectional lstm cell toward hate speech prediction," *IEEE Access*, vol. 11, pp. 16 801–16 811, 2023.