International Research Journal of Mathematics, Engineering and IT

**ISSN: (2348-9766)**
**Association of Academic Researchers and Faculties (AARF)**
**Impact Factor**- 5.489Volume 11, Issue 3, March 2024
**Website**- www.aarf.asia, **Email** : editor@aarf.asia , editoraarf@gmail.com

# Exploring LSTM Networks and Word Embeddings for Hate Speech and Offensive Language Detection on Social Media

Renu Kumari, *Research Scholar*, *Pataliputra University*

Vijay Kumar

*College of Commerce Arts and Science Patliputra University, Patna, Bihar, India-800020*

## Abstract

The detection of hate speech on social media platforms like Twitter is a critical area of research due to its potential societal impact. In this study, we explore the application of deep learning techniques for hate speech detection on Twitter data. Our approach involves several experiments aimed at identifying hate or offensive words, a crucial step in distinguishing between hate speech and non-hate speech content. Leveraging the Twitter API, we obtained a dataset comprising 11,325 annotated tweets categorized into 'Sexism', 'Racism', and 'None' classes. Through the use of deep learning, we trained our models to learn abstract feature representations from the input data. Specifically, our deep learning models are designed to automatically extract meaningful features from the text data, enabling effective discrimination between hate speech and normal content. We evaluated the performance of our models on the Twitter dataset to assess their efficacy in hate speech detection. Our findings demonstrate the potential of deep learning techniques in accurately identifying hate speech on Twitter, thereby contributing to the development of more robust and effective mechanisms for combating online hate speech and fostering a safer online environment. *Keywords:* Hate speech, Machine learning, Online Social network, Social Media

International Research Journal of Mathematics, Engineering and IT

ISSN: (2348-9766)
Association of Academic Researchers and Faculties (AARF)
Impact Factor- 5.489Volume 11, Issue 3, March 2024
Website- www.aarf.asia, Email : editor@aarf.asia , editoraarf@gmail.com

## Introduction

Social media platforms like Twitter have become ubiquitous channels for communication, enabling users to express opinions, share information, and engage in discussions on diverse topics. However, alongside the benefits of these platforms, there exists a dark side characterized by the pro-liferation of hate speech, which poses significant challenges to societal harmony, online discourse, and individual well-being. Hate speech, defined as communication that disparages individuals or groups based on characteristics such as race, ethnicity, religion, gender, sexual orientation, disability, or other attributes, has increasingly garnered attention due to its potential to incite violence, perpetuate discrimination, and foster social division [1].

Despite the advancements in hate speech detection techniques, several challenges persist in effectively identifying and addressing hate speech on social media platforms like Twitter. Firstly, the rapid evolution of language and the emergence of new forms of hate speech pose significant challenges to static detection models, necessitating constant adaptation and updates [2]. Secondly, the contextual ambiguity inherent in many tweets can lead to misinterpretation, making it challenging for automated systems to accurately discern between hate speech and other forms of expression [3]. Additionally, the prevalence of implicit and coded language used to circumvent detection algorithms further complicates the task of identifying hate speech [4]. Furthermore, the diverse cultural and linguistic backgrounds of users contribute to the variability in hate speech ex-pressions, requiring detection models to be sensitive to linguistic nuances and cultural contexts [5].

Addressing these challenges requires innovative approaches that leverage advanced deep learning techniques like LSTM networks to effectively navigate the complexities of hate speech detection in Twitter data.

Detecting and mitigating hate speech on social media platforms has thus emerged as a critical area of research and intervention. Traditional methods of hate speech detection often rely on lexicon-based approaches or handcrafted features, which struggle to capture the

nuanced and evolving nature of hate speech language [6]. In contrast, deep learning techniques, particularly Long Short-Term Memory (LSTM) networks, have shown promise in automatically learning complex patterns in sequential data, making them well-suited for hate speech detection tasks on platforms like Twitter [7].

This paper aims to investigate the effectiveness of deep learning-based LSTM models in de- testing hate speech within Twitter data. By leveraging the temporal dynamics and contextual information present in tweets, we seek to develop a robust and scalable approach to identify and classify instances of hate speech accurately. Through empirical evaluation and comparison with existing methods, we aim to contribute to the advancement of hate speech detection techniques, there by facilitating the development of more effective strategies for combating online hate and promoting healthier digital environments.

For our methodology, we employ Long Short-Term Memory (LSTM) networks, a type of recur- rent neural network (RNN), renowned for their capability to capture long-range dependencies and temporal dynamics in sequential data such as text. We preprocess the Twitter data by tokenizing and embedding the tweets into numerical vectors, preserving the semantic meaning of words while encoding them into a format suitable for deep learning models [8]. We then construct and LSTM-based architecture, comprising multiple layers of LSTM units followed by fully connected layers for classification. The model is trained on labeled Twitter data annotated for hate speech presence, utilizing techniques such as mini-batch stochastic gradient descent and backpropagation to optimize the model parameters and minimize the classification error [7]. To evaluate the performance of our LSTM-based hate speech detection model, we employ standard metrics such as, precision, recall, and F1-score, comparing its effectiveness against baseline models and existing state-of-the-art methods [6].

## 2. Related Work

Hate speech, defined as communication that disparages individuals or groups based on characteristics such as race, ethnicity, religion, gender, sexual orientation, disability, or other attributes, has increasingly permeated online platforms, posing significant challenges to societal harmony, online discourse, and individual well-being. Social media platforms like

ISSN: (2348-9766)
Association of Academic Researchers and Faculties (AARF)
Impact Factor- 5.489Volume 11, Issue 3, March 2024
Website- www.aarf.asia, Email : editor@aarf.asia , editoraarf@gmail.com

Twitter, with their widespread accessibility and ease of communication, have become hotbeds for the propagation of hate speech. Detecting and mitigating hate speech on such platforms has thus emerged as a critical area of research and intervention. This literature survey aims to explore significant contributions in the field of hate speech detection, particularly focusing on studies related to Twitter data. By examining various methodologies, evaluations, and task-specific investigations, we aim to provide insights into the current state of research and identify emerging trends and challenges in hate speech detection on social media platforms.

Davidson et al. [1] proposed an automated approach for hate speech detection on social media, particularly focusing on the challenges posed by offensive language. Their study emphasized the need for sophisticated algorithms to effectively identify and mitigate hate speech online. Waseem and Hovy [2] conducted a detailed investigation into predictive features for hate speech detection specifically on Twitter. Their work provided insights into distinguishing between hate speech directed at individuals and hate speech involving hate symbols, contributing to a nuanced understanding of hate speech dynamics. Fortuna et al. [6] conducted a comprehensive survey of automatic hate speech detection methods in text. Their survey covered a wide range of techniques, including lexicon-based approaches, machine learning models, and deep learning architectures, providing a valuable resource for researchers and practitioners in the field. Burnap and Williams [2] analyzed hate speech on Twitter using machine learning and statistical modeling techniques to inform policy-making decisions. Their study highlighted the importance of leveraging computational methods to understand and address the proliferation of hate speech on social media platforms. Founta et al. [9] conducted a large-scale crowdsourcing effort to characterize abusive behavior on Twitter. Their work shed light on the prevalence and nature of abusive behavior online, contributing valuable insights for hate speech detection research and online safety initiatives.

Fersini et al. [10] provided an overview of the Evalita 2018 Hate Speech Detection Task,which aimed to advance hate speech detection techniques for the Italian language. Their work highlighted the importance of task-specific evaluations in advancing the state-of-the-art in hate speech detection. Chen et al. [11] developed a deep learning-based model

ISSN: (2348-9766)
Association of Academic Researchers and Faculties (AARF)
Impact Factor- 5.489Volume 11, Issue 3, March 2024
Website- www.aarf.asia, Email : editor@aarf.asia , editoraarf@gmail.com

for detecting offensive language on social media platforms, including Twitter, to enhance adolescent online safety. Their study demonstrated the effectiveness of advanced machine learning techniques in addressing the challenges of detecting offensive language online. Wulczyn et al. [12] explored personal attacks on Wikipedia using machine learning techniques. Their work provided valuable insights into the challenges and opportunities for detecting similar behavior on social media platforms like Twitter,highlighting the importance of understanding context and social dynamics in hate speech detection.

Schmidt and Wiegand [4] conducted a survey on hate speech detection using natural language processing techniques, with a focus on approaches specifically designed for Twitter data. Their study, reviewed various methods and highlighted the need for robust and context-aware algorithms for hate speech detection on social media platforms. Iwendi et al. [13] proposed a deep learning approach based on Long Short-Term Memory (LSTM) networks for detecting cyberbullying on Twitter. Their study, presented at the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), demonstrated promising results in identifying abusive behavior online, paving the way for more advanced techniques in hate speech detection. This detailed literature survey provides a comprehensive overview of significant research contributions in the domain of hate speech detection, encompassing various methodologies, evaluations, and task-specific investigations.

## 3. Dataset Description

The dataset utilized in this study comprises publicly available English tweets collected from the social media platform Twitter. Twitter's API was employed to retrieve individual tweets along with user details based on the provided Tweet ID. The dataset is sourced from the GitHub repository, accessible at: [1]. Each tweet is associated with a unique Tweet ID and categorized into one of three classes: Sexism, Racism, and None. 'Sexism' denotes tweets containing sexist content, 'Racism' indicates tweets containing racist content, and 'None' represents tweets categorized as non-hate speech. There are total of 11325 tweets are available in which 2, 988 tweets aare categorized as 'sexism', 20 tweets as

**ISSN: (2348-9766)**
**Association of Academic Researchers and Faculties (AARF)**
**Impact Factor**- 5.489Volume 11, Issue 3, March 2024
**Website**- www.aarf.asia, **Email** : editor@aarf.asia , editoraarf@gmail.com

'racism' and 8317 tweets as normal which means non-hate speech.

## 4. Proposed Method

In this section, we present our proposed methodology for hate speech detection on Twitter using Long Short-Term Memory (LSTM) networks and Word2Vec. The work flow of the proposed model is well described by the figure 1.

### *4.1. Data Preprocessing*

In this phase, we have checked that is there any missing values (null values) in the dataset or not and we found that there are no missing values present in the dataset. First we have removed any characters that are not letters (i.e., non-alphabetic characters) from the raw tweet using the regular expression and replaces them with spaces. All the letters in the tweets are converted to lowercase and then splitted into individual words, creating a list of words. Stopwords like "and" "the," "is" are often removed from text data during text preprocessing because they typically do not carry significant meaning. Before feeding the Twitter data into the LSTM model, we preprocess the text by tokenizing the tweets and converting each token into a numerical representation. This

[1]https://github.com/srishb28/Hate-Speech-Detection-on-Twitter-Data/blob/master

International Research Journal of Mathematics, Engineering and IT

ISSN: (2348-9766)
Association of Academic Researchers and Faculties (AARF)
Impact Factor- 5.489Volume 11, Issue 3, March 2024
Website- www.aarf.asia, Email : editor@aarf.asia , editoraarf@gmail.com

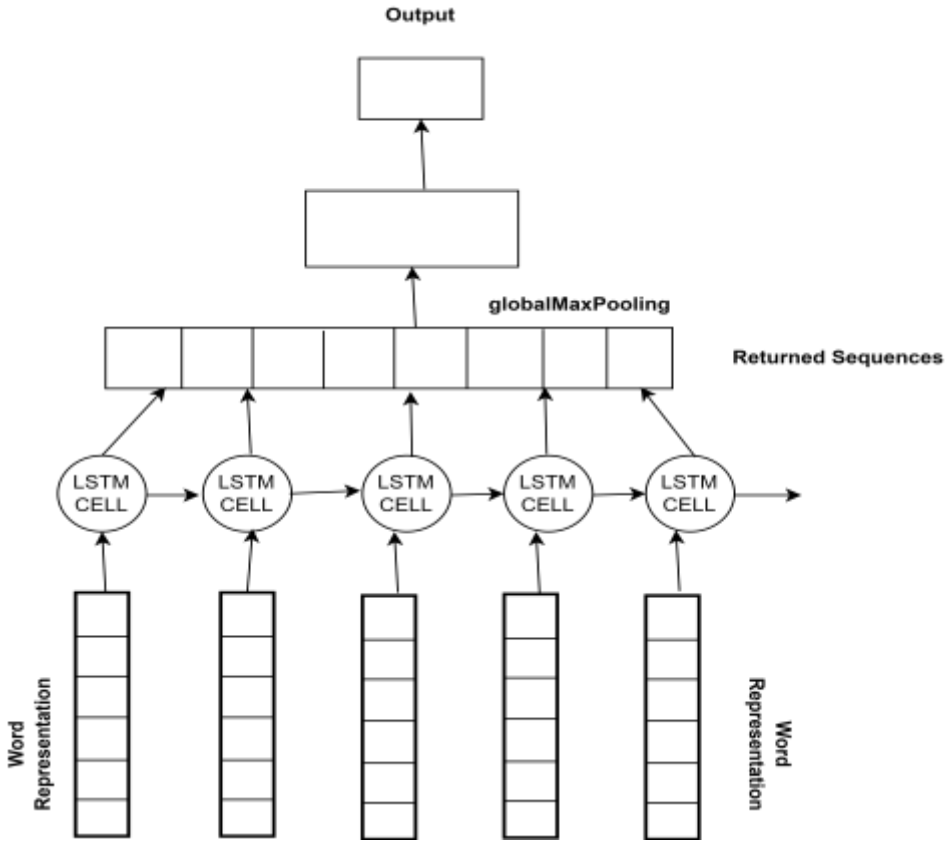


Figure 1: LSTM Networks and Word Embeddings based Hate Speech and Offensive Language Detection Model

is typically achieved using word embedding technique called Word2Vec, which map each word to a low dimensional dense vector in a continuous space.

### *4.2. Data Splitting*

The dataset is partitioned into training and testing sets, with 11, 000 tweets allocated to the training set and the remaining tweets are assigned to the test set.

### *4.3. LSTM Architecture*

The LSTM model consists of multiple LSTM layers followed by a fully connected layer for classification. Let $X = \{x_1, x_2, \ldots, x_n\}$ represent the input sequence of Twitter messages, where each $x_i$ is a word or token in the message. Each word is represented as a word embedding vector $\mathbf{w}_i$. The hidden state of the LSTM at time step $t$, denoted as

International Research Journal of Mathematics, Engineering and IT

ISSN: (2348-9766)
Association of Academic Researchers and Faculties (AARF)
Impact Factor- 5.489Volume 11, Issue 3, March 2024
Website- www.aarf.asia, Email : editor@aarf.asia , editoraarf@gmail.com

$h_t$, is computed recursively by usingthe equation 1.

$$\mathbf{h}_t, \mathbf{c}_t = \text{LSTM}(\mathbf{h}_{t-1}, \mathbf{w}_t) \quad (1)$$

where $\mathbf{h}_{t-1}$ is the previous hidden state, $\mathbf{w}_t$ is the input word embedding at time step $t$, and $\mathbf{c}_t$ is the cell state. The LSTM cell computes three gates: the input gate $i_t$, the forget gate $f_t$, and the output gate $o_t$, as well as the cell state update $\tilde{\mathbf{c}}_t$. All the three gates: input gate $i_t$, forget gate $f_t$, and output gate $o_t$, as well as the cell state can be described by the equations 2, 3, 4, 5 and 6 respectively.

$$i_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{w}_t] + \mathbf{b}_i) \quad (2)$$

$$f_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{w}_t] + \mathbf{b}_f) \quad (3)$$

$$o_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{w}_t] + \mathbf{b}_o) \quad (4)$$

$$\tilde{\mathbf{c}}_t = \tanh(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{w}_t] + \mathbf{b}_c) \quad (5)$$

$$\mathbf{c}_t = f_t \odot \mathbf{c}_{t-1} + i_t \odot \tilde{\mathbf{c}}_t \quad (6)$$

$$\mathbf{h}_t = o_t \odot \tanh(\mathbf{c}_t) \quad (7)$$

where $\sigma$ denotes the sigmoid function, $\odot$ represents element-wise multiplication, $[\mathbf{h}_{t-1}, \mathbf{w}_t]$ denotes concatenation, and $\mathbf{W}$ and $\mathbf{b}$ are weight matrices and bias vectors, respectively.

### *4.4. ClassiFIcation*

The final hidden state $\mathbf{h}_n$ of the last LSTM layer is passed through a fully connected layer with softmax activation to obtain the probability distribution over classes and it can be observed from the equation 8.

$$\hat{y} = \text{softmax}(\mathbf{W}\mathbf{h}_n + \mathbf{b}) \quad (8)$$

### *4.5. Training*

The model is trained using categorical cross-entropy loss, defined by equation 9.

$$\text{Loss} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (9)$$

International Research Journal of Mathematics, Engineering and IT

ISSN: (2348-9766)
Association of Academic Researchers and Faculties (AARF)
Impact Factor- 5.489Volume 11, Issue 3, March 2024
Website- www.aarf.asia, Email : editor@aarf.asia , editoraarf@gmail.com

where $C$ is the number of classes, $y_i$ is the ground truth label for class $i$, and $\hat{y}_i$ is the predicted probability for class $i$. The other parameters of the model, including the weights and biases of the LSTM layers and the fully connected layer, are updated using backpropagation through time (BPTT) and stochastic gradient descent (SGD)optimization

## 5. Result Analysis

The performance of the LSTM-based hate speech detection model is evaluated using standard metrics such as precision, recall, and F1-score on a held-out test dataset. Additionally, we conduct experiments to compare the proposed LSTM model with baseline models and state-of-the-art approaches.

### *5.1. Evaluation Metrics*

The evaluation metric utilized throughout the experiment is designed to account for positive and negative influences on precision, recall, and F1-scores. Positive speech denotes hate speech, while negative speech refers to non-hateful speech. Precision measures the proportion of correctly identified positive cases out of all cases predicted as positive. Recall, also known as sensitivity, measures the proportion of correctly identified positive cases out of all actual positive cases. F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics.Precision, Recall and F1 Score are calculated as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (1)$$

$$\text{F1-Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (4)$$

These measures collectively contribute to the comprehensive evaluation of our model's performance and the validity of our results.

### *5.2. Layer Structure of the DeFIned Model*

We have provided the construction and training of a deep learning based hate speech detection model. The input layer is defined with a shape of 53 indicating that each input instance has 53features. The embedding layer converts input indices into dense vectors of dimension 200, using embeddings provided in the embedding matrix. The LSTM layer consists of 50 units, with a return sequence set to True to return the full sequence of outputs. Dropout and recurrent dropout are applied to the LSTM layer to prevent overfitting. GlobalMaxPool1D layer is used to reduce the dimensionality of the output from the LSTM layer. Dense layers with 50 units are added with LeakyReLU activation function. Dropout layer is added to prevent overfitting. The output layer consists of 3 units with softmax activation. The model is compiled with categorical crossentropy loss function, and Adam optimizer. The model is trained for 25 epochs with a batch size of 32 and a validation split of 0.1.

### *5.3. Results*

The table 1 and Figure 2 present performance metrics, including precision, recall, and F1-score, for five different models: SVM, LR, NB, KNN and LSTM. SVM achieves moderate precision and recall but a relatively lower F1-score compared to other models. It seems to have balanced precision and recall. LR shows a higher precision compared to SVM but slightly lower recall. Its F1-score is moderate. NB performs similarly to SVM in terms of precision, recall, and F1-score. KNN demonstrates a good balance between precision and recall, resulting in a relatively high F1-score. LSTM outperforms other models in terms of precision and F1-score, indicating its ability to make accurate positive predictions. However, its recall is slightly lower compared to KNN. The LSTM

International Research Journal of Mathematics, Engineering and IT

ISSN: (2348-9766)
Association of Academic Researchers and Faculties (AARF)
Impact Factor- 5.489Volume 11, Issue 3, March 2024
Website- www.aarf.asia, Email : editor@aarf.asia , editoraarf@gmail.com

Table 1: Performance comparision of the proposed model with various sate-of-the arts

| Models | Precision | Recall | F1-score |
|---|---|---|---|
| SVM | 0.56 | 0.75 | 0.64 |
| LR | 0.74 | 0.67 | 0.69 |
| NB | 0.56 | 0.75 | 0.64 |
| KNN | 0.73 | 0.76 | 0.73 |
| LSTM | **0.88** | 0.71 | **0.8** |

model outperforms all the models in all three metrics, demonstrating higher precision, recall, and F1-score values, indicating better performance in classifying instances of interest.

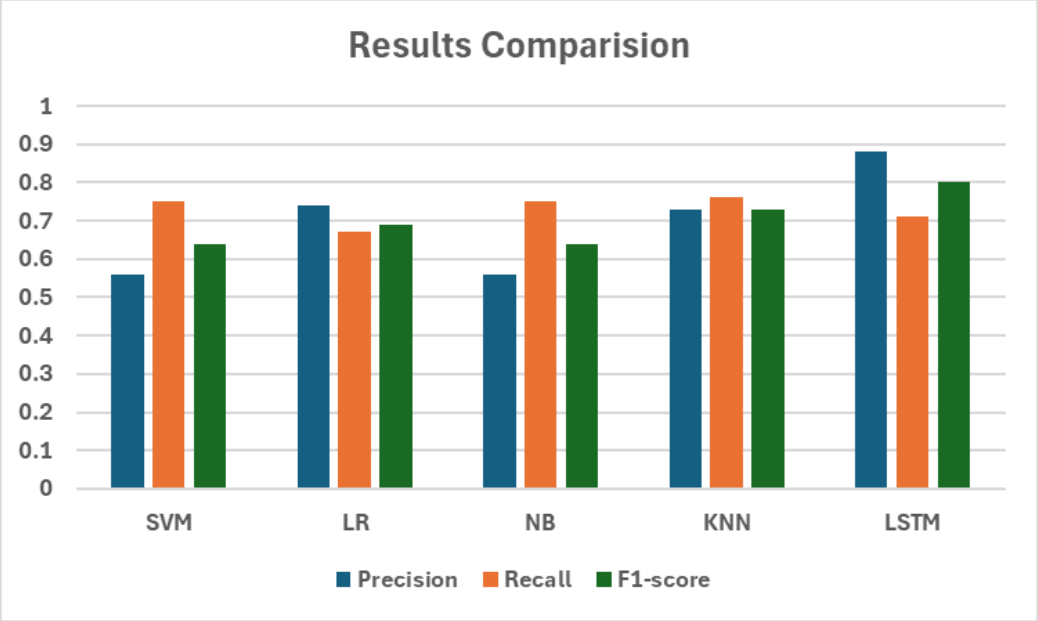


Figure 2: LSTM Networks and Word Embeddings based Hate Speech and Offensive Language Detection Model

In summary, each model has its own trade-offs between precision, recall, and F1-score. SVM and NB have lower precision and F1-score, LR shows better precision but slightly lower recall, KNN achieves a good balance between precision and recall, and LSTM has the highest precision but relatively lower recall compared to KNN.

## 6. Conclusion and Future Work

Hate speech detection has emerged as a critical issue, and implementing an automated system to detect such content is seen as a viable solution to address this pressing concern. Machine learning algorithms have been proposed as effective tools for identifying hate speech and offensive language across various online platforms susceptible to such content. The primary aim was to develop an automated method to mitigate social animosity prevalent in social media and online communities. The proposed LSTM model demonstrated superior performance compared to existing approaches, achieving a commendable accuracy of 92.1% when evaluated on test data. To address the challenge of identifying offensive language that may not contain explicit hostile phrases, additional examples were collected. However, a limitation of the model is its inability to consider the context of negative words within a phrase, which could be improved by integrating linguistic features. Two notable constraints of the study include the model's lack of real-time prediction accuracy and its inability to assess the intensity of hate speech messages. Consequently, the proposed machine learning model could also be employed to predict the severity of hateful messages. Hate speech detection poses significant challenges due to the diverse nature of content found in social media posts, which often include images and code-mixed languages. To enhance the robustness of the algorithm, future research on hate speech detection should incorporate multilingual examples and consider different modalities of social media content.

**References**

[1] T. Davidson, D. Warmsley, M. Macy, 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, 2017, pp. 512–515.

[2] Z. Waseem, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[3] P. Burnap, M. L. Williams, Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & internet 7 (2) (2015)

223–242.

[4] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the fifth international workshop on natural language processing for social media, 2017, pp. 1–10.

[5] J. C. Pereira-Kohatsu, L. Quijano-S ánchez, F. Liberatore, M. Camacho-Collados, Detecting and monitoring hate speech in twitter, Sensors 19 (21) (2019) 4654.

[6] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (4) (2018) 1–30.

[7] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997)1735–1780.

[8] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781.

[9] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A.Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of twitter abusive behavior, in: Proceedings of the international AAAI conference on web and social media,Vol. 12, 2018.

[10] E. Fersini, P. Rosso, M. Anzovino, et al., Overview of the task on automatic misogyny identification at ibereval 2018., Ibereval@ sepln 2150 (2018) 214–228.

**ISSN: (2348-9766)**
**Association of Academic Researchers and Faculties (AARF)**
**Impact Factor**- 5.489Volume 11, Issue 3, March 2024
**Website**- www.aarf.asia, **Email** : editor@aarf.asia , editoraarf@gmail.com

[11] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, IEEE, 2012, pp. 71–80.

[12] E. Wulczyn, N. Thain, L. Dixon, Ex machina: Personal attacks seen at scale, in: Proceedings of the 26th international conference on world wide web, 2017, pp. 1391–1399.

[13] C. Iwendi, G. Srivastava, S. Khan, P. K. R. Maddikunta, Cyberbullying detection solutions based on deep learning architectures, Multimedia Systems 29 (3) (2023) 1839–1852.