Hate Speech Recognition System through NLP and Deep Learning

Sagar Mujumale[1], Prof. Nagaraju Bogiri[2]

[1,2] Computer Department, KJCOEMR, Pune, India

***Abstract*–Hate speech is connected to racial prejudice, and there is evidence that hate crimes are on the rise. It has grown in popularity as a result of the rise of online social media where most hate speech is concentrated. Several government-sponsored remedies are being undertaken as the problem of racist speech acquires traction. In recent years, there has been a fast expansion of information or knowledge, which has been driven by the internet paradigm. The quick expansion has resulted in the realization of a wide range of distinct and one-of-a-kind implementations. These services have rapidly expanded and are providing increasing convenience in all aspects of life, including sociability. Socialization has been moved online through the usage of online social networks, which have grown in popularity in recent years. Every day, new people join these online platforms, significantly boosting their user base while also increasing the incidents of hate speech. Therefore, this research article elaborates on an effective approach for the purpose of achieving effective hate speech recognition through the use of Natural Language Processing approaches such as TF-IDF, Entropy Estimation along with Fuzzy Artificial Neural Networks and Decision Making. The experimentations have been conducted to attain the performance of the approach which has resulted in highly positive results.**

Keywords— *Hate Speech Recognition, Fuzzy ANN, Entropy Estimation, Natural Language Processing, Term Frequency and Inverse Document Frequency.*

I. INTRODUCTION

Internet has grown in popularity as an excellent means to convey one's feelings and emotions. However, under the pretext of free expression, the increasing use of social media has resulted in the spread of hate propaganda. Despite the fact that social media is extremely quick, open, free, and simple to use, it is also quite vulnerable due to its rapid growth. It is used by miscreants to promote various types of bigotry or prejudice statements directed towards another community. Hate speech is described as discourse that may be damaging to a person's or group's feelings and may inspire violence or a lack of compassion, as well as irrational and inhuman behavior.

These socializing services allow users to communicate and socialize with their friends and followers by exchanging text messages and material such as photographs, videos, and so on. The social network concept allows users to communicate with their followers and contribute their opinions, which are relayed to them. More than any other sort of material, social networking sites and tweeting services attract Internet users. Facebook, Instagram, and Twitter are growing increasingly popular among people of different ages, races, and interests. Their material is continually expanding, making them an intriguing example of so-called big data. Big data has aroused the interest of academics who, among other things, want to automate the study of people's thoughts and the organization or dispersion of users in companies.

Social media is frequently used to disseminate a wide range of content. People commonly use social media to express their opinions and ideas. While these platforms allow users to explore and share their ideas, the sheer amount of postings, comments, and conversations makes maintaining quality control challenging [1]. Furthermore, due to the diversity of origins, cultures, and beliefs, many people use aggressive and harsh language while interacting with persons of different nationalities.

Interaction is one of the most important aspects of a person's daily life. Humans and other primates have a universal need for sociability, which has been seen and extensively documented. Individual communication is nearly totally responsible for socialization. The primary goal of socializing is to promote healthy dialogue, which allows for successful mood enhancement and the formation of social bonds. A lack of communication or sociability may be immensely damaging to an individual's general welfare, leading to a variety of mental health issues.

There are several approaches for socializing or communicating with one another, but one of the most frequent is through the use of speech or language. This is one of the most powerful and popular types of socializing found all across the world. This may be seen in the wide variety of languages that have emerged and are now used all across the world. Over the history of human evolution, there have been clever storytelling and other modes of communication that have culminated in the creation of language that we see and use today including the hate speech which is highly difficult to detect automatically [2].

The second section of this research article is a literature review. The proposed approach is described in section 3, and the acquired findings are carefully assessed in part 4. This study article is finalized in the section 5 including the extent of the future improvements.

II. LITERATURE SURVEY

Pradeep Kumar Roy [3] addresses the issue of hate speech detection on Twitter using a deep convolutional neural network. The authors have utilized approaches such as KNN, GB, DT, SVM, NB, RF, and LR which are machine learning classifiers. These classifiers have been utilized for the purpose of achieving the identification of the hate speech content through the TF-IDF values retrieved from the tweets. The classifiers were then compared with the one another on the basis of their classification accuracy. The classification accuracy with these classifiers has been on par with the conventional CNN methodology. The proposed DCNN approach achieves the optimal solution and an improved accuracy over the classifiers that have been tested.

Flor Miriam Plaza-Del-Arco [4] states that the problem of hate speech is one of the most problematic occurrences that have been a significant challenge for the social media networks and other web based platforms. Investigations on two benchmark crops show that their proposed technique outperforms an STLBETO model and yields state-of-the-art results. The proposed model's outcomes, as well as a complete knowledge acquisition research from SA, show that polarization and emotional classification methods help the MTL model recognize HS by leveraging emotional information. The relationship between emotional knowledge and HS opens the door to new methods to constructing NLP systems in other disciplines where polarity and emotion may be important.

Ashwin Geet d'Sa [5] investigated the multiclass categorization of hate speech using embedding vector representations of words and DNNs. The classification was performed on Twitter data that used a three-class classification scheme: hate, offensive, and neither. They proposed feature-based and fine-tuning strategies for hate speech classification. The feature-based method generates a series of word embeddings as source for the classifiers. As word embeddings, they investigated fastText embedding and BERT embedding. Within the framework of a feature-based approach, the capabilities of these two types of embeddings are almost comparable.

C. Baydogan [6] proposes two unique optimization-based strategies for tackling the HSD challenge in social networking websites. For the first time in the study, the most recent optimization techniques, ALO and MFO, were used to address the HSD problem. Researchers used 8 different supervised data mining algorithms, SSO, and cutting-edge TSA to monitor the effectiveness of the recommended metaheuristic-based approaches. The pre-processing stage was completed using NLP techniques for the given real-world HSD situations. The BoW+TF+Word2Vec approaches were used to extract features. Then, in order to address HSD concerns, twelve different algorithms competed. With the exception of one dataset, the modified ALO algorithm produced the highest accuracy, specificity, responsiveness, and f-score numbers in the study.

Y. Zhou [7] presented the principles of three different types of text classification methods, ELMo, BERT, and CNN, and used them to detect hate speech. He then enhanced the efficiency by blending from multiple viewpoints: blending of ELMo, BERT, and CNN classification techniques, and merging of 3 CNN classifiers with varying parameters. The findings indicated that unification synthesis can help identify hate speech.

M. Mozafari [8] investigated the feasibility of a meta-learning approach as a viable strategy of few-shot acquisition in cross-lingual hateful speech and inappropriate language detection tasks for the first time. To that end, the authors created two evaluation metrics for cross-lingual hateful speech and inappropriate language classification techniques by combining a variety of publicly accessible datasets including hate and controversial material from a variety of languages. The authors employed a meta-learning strategy based on multi-threading and metric-based techniques to train the model that can generalization quickly to a new language with a small amount of classification model (k examples per class) (MAML and Proto-MAML). The results reveal that meta learning-dependent models outperform transferable learning-based algorithms in the majority of circumstances, with Proto-MAML being just the best model for recognizing hostile or offensive language with a little quantity of labeled data.

M. Z. Ali [9] created a comprehensive data collection by gathering Urdu language tweeting and getting trained linguists analyze them on aspect and mood levels. There is currently no data collection of annotated Urdu inciting hatred with element and emotion degrees. The authors applied cutting-edge methodologies to overcome the three most prominent problems in deep learning sentiment classification, including sparsity, complexity, and class skew, and saw an improvement in performance well over model generated. Two machine learning techniques were used to train the classifier: SVM as well as Multilayer perceptron Nave Bayes. To minimize sparsity, the authors used dynamic stop words filtering, a variable global feature selection strategy, and artificial minority frame interpolation to decrease class imbalance.

O. Oriola [10] built an English collection of South African tweets in applied to measure objectionable and hateful speech. The collection was transcribed by multilingual transcribers because the tweets featured a variety of indications from Southern African languages. Four distinct extracted features and their permutations were extracted from the tweets after tokenization and preprocessing. Researchers used three types of improved machine learning models to classify tweets as hate speech, offensive speech, or free speech: hyper-parameter optimization, ensemble, and multi-tier meta-learning on different machine learning algorithms such as Regression Model, Support Vector Machine, Multilayer Perceptron, and Random Forest.

H. S. Alatawi [11] investigated network and agnostic word encoding using deep learning. According to the data, this strategy is helpful in suppressing white nationalist hateful speech. The BERT approach has also shown to be the most up-to-date answer for this problem. The trial results show that BERT outperforms domain-specific method by 4 points; nevertheless, the domain-specific technique can distinguish purposely misspelled terms and common slang from the hate movement, but the BERT modeling cannot even though it is learned on Wikipedia and culture. Some dataset in the investigations are imbalanced in order to mimic actual information, while others are balancing in order to assess the performance of the model under perfect circumstances.

L. H. Son [12] developed sAtt-BLSTM convNet, a mix of soft attention-dependent bidirectional long short-term memory (sAtt-BLSTM) as well as convolution neural network (convNet), to identify sarcasm in short texts (tweets). Semantic word embeddings as well as pragmatic auxiliary characteristics were used to train the network. In comparison to the baseline approaches, the suggested model has the highest classification accuracy across both datasets. The use of mash-up languages and novel vocabulary with complicated structures increases the challenge of automatic sarcasm detection and highlights several unresolved issues.

V. -I. Ilie [13] talks about his work on ContextAware Misinformation Identification Employing Deep Learning Implementations. They employ two text preprocessing pipelines (Lemma and Aggressive Text Preprocessing) for multi-class classification, 3 context-aware phrase embeddings, and ten Machine Learning. Context-aware extracted features are either pre-trained or custom generated on the database. They propose a processing and categorization pipeline based on their findings. The experimental validation dataset consists of several news articles classified as true or false.

H. Watanabe [14] presented a pattern-based approach for detecting hate speech on Twitter. The authors provide a set of settings to maximize pattern collection by pragmatically dynamic and ever changing from the training set. They also offer a system for detecting hate speech in which words and phrases pragmatically signifying hatred and offense are accumulated and mixed with themes and other sentiment-based qualities. The recommended unigram and trend collections will be utilized as pre-built dictionaries for subsequent hate speech recognition investigations. They divide comments into three main categories to distinguish between aggressive and just offensive tweets.

III PROPOSED METHODOLOGY

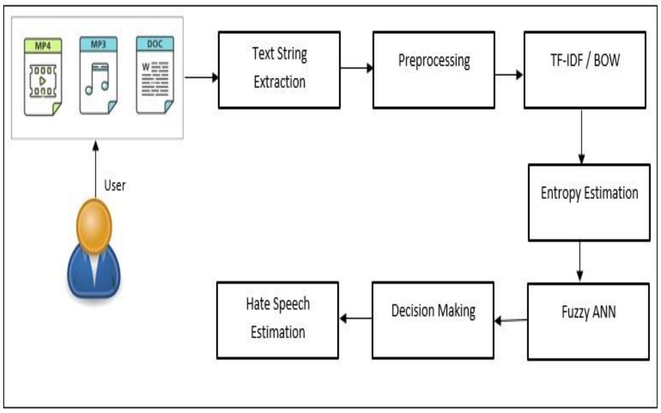


Figure 1: System Overview Diagram

The proposed approach for the purpose of achieving the hate speech recognition through the use of Fuzzy Artificial Neural Network has been described in the steps give below.

Step 1: Data Collection and Preprocessing – This is the initial step of the methodology where an excel sheet is provided as an input to the proposed approach which contains tweets suspected of propagating hate speech. This tweet Dataset is downloaded from the URL: https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv. The audio files with the hate speech if any, are provided to a python code that effectively extracts the speech text from the audio and adds it to the excel sheet.

The excel sheet is provided to the system that is a java code. The JXL library is being utilized to interface the workbook format file to the java code effectively. The contents of the excel file are being converted into a string format and then stored in the form of a list which is then provided to the next step of the approach for the purpose of preprocessing.

Step 2: Preprocessing – This is one of the most essential steps in the approach due to the fact that this step can considerably improve the execution performance effectively. The extracted text string from the previous step with suspected hate speech is taken as an input in this step of the approach. This step identifies and removes any redundancies or contradictions in the string to clean or condition it. This process considerably reduces the errors and any kind of flaw in the string that can cause a bottleneck in the later sections of this approach.

The preprocessing is achieved by a number of steps elaborated below.

Special Symbol removal – The string used as input in the structure of a workbook would contain a variety of special characters that users would use to write the content using correct syntax. Among these special characters are #, !, @,?, and Such special symbols can dramatically increase processing time, hence raising the approach's temporal complexity. As a result, these special symbols can indeed be removed without consequence.

Stemming – This step is one of the most successful because it uses the input string to seek for words with postfixes in their end such as ion, ing. These postfixes are just extensions of the core word. These prefixes are unnecessary and superfluous in the input text since they provide no further information. As a result, these words must be reduced to the base word, such as going, which will be changed into go without affecting the string's underlying meaning.

Stop Word Removal – Stop words are words in the English language that serve as a conjunction or link between two elements of a phrase. These words are extremely important in spoken English because they construct a model to the phrase that a listener can efficiently follow. These words are mainly cosmetic in design and provide no further semantic value to the phrases. As a result, terms like from, and, is, the, and so on may be simply deleted and discarded from the input text without affecting the semantic information of the string, substantially lowering the time required for system execution significantly.

Step 3: Term Frequency & Inverse Document Frequency – In this stage of the suggested technique, which is one of the most significant, the input string is examined using the TF-IDF model. This TF-IDF approach identifies the importance of the words in the string by assessing the term frequency and inverse document frequency. This step can be quantitatively expressed using the accompanying equation 1.

$$\mathrm{TF}-\mathrm{IDF} = TF\ X\ Log\,\frac{Number\ of\ Documents}{Number\ of\ Documents\ Containing\ Word\ W} \text{ ---- (1)}$$

Term frequency is obtained by calculating the frequency of each term in a given comment or hate speech. The Term Achieved the prevalence of a term is calculated by multiplying the logarithmic ratio of the total number of documents by the number of documents containing a specific word W.

The words with the highest TF-IDF scores are considered to have a substantial impact on the sarcastic pattern's creation. As a consequence, these words have been categorized in the created framework for future use. The TF-IDF technique may be defined using the algorithm 1 shown below.

ALGORITHM 1: TF-IDF Estimation

```
Start
Read the Preprocessed string
Divide string into words using space and store in a vector V
    For i =0 to N (Where N is the length of V)
        W= V[i]
        Count W for the respective string as TF
        Count W for the all other input strings that is DF
        IDF= log (DF)
        TF-IDF= TF* IDF
    End For
Stop
```

Step 4: Bag of Words: Alongside TF-IDF, this is yet another essential component of Natural Language Processing that's also implemented in the proposed approach. The bag of model likes to examine the existence of hate speech in the supplied input text in the preprocessed string. The Bag of Terms is a set of words that have been expressly designated as hateful and used by the individuals engaging in hate speech.

The list previously obtained in the earlier step is being used as an input in this step of the approach. This input list contains the words in the first column and the TF-IDF scores in another. This list of words is correlated with the bag of Words stored in the database. If the word in the list is correlated with the Bag of Words, then the score 1 is appended at the end of the list, if the word is not found in the BoW then a score of 0 is appended. This is done for all of the words in the list and an updated list with the BoW score column is provided to the next step for entropy evaluation.

Step 3: Entropy Estimation – The information gain values of the obtained word characteristics for the input string will have to be analyzed. This stage of the technique uses the list obtained from the previous stage as an input. The entropy of the BOW words is calculated using Shannon information gain, which is provided in equation 2 below.

$$\mathrm{E} = -\frac{a}{c}\log\frac{a}{c} - \frac{b}{c}\log\frac{b}{c} \_\_\_\_ (2)$$

Where,

a= matched word count
c= total number of tweets
b= c-a
E = Entropy Gain factor

The words are retrieved for all of the BoW items, and the number of comparable words is tallied using entropy, which would be referenced to as the Information gain score. This score is calculated using the Shannon information gain formula mentioned before. The entropy estimates obtained are then added to the list and sent on to the next phase for the further evaluation.

Step 5: Fuzzy Artificial Neural Network – This is among the most fundamental components in the proposed approach, in which a double-dimensional list of features obtained previously is used to efficiently used to generate the neurons for the Artificial Neural Networks. The hidden layer and the output layer values are evaluated using relevant bias weights and target values as shown in the equation 3 given below. The equation 4 is being used for the purpose of hidden layer values estimation. The equation 5 provides the formula for the activation function called ReLU being used in the Fuzzy ANN.

$$\mathrm{T}=\left(\sum_{k=0}^{n} AT * W\right) + B \_\_\_ (3)$$

$$\mathrm{H_{LV}} = 2 \left(\frac{1}{(1+\exp(-T))} \; X \; 2T \right) -1 \_\_\_\_\_ (4)$$

Where,
n- Number of attributes
A_T- Attribute Values
W- Random Weight
B- Bias Weight
H_{LV} – Hidden Layer Value

$$f(x) = max(0, x) \_\_ (5)$$

The difference between the highest and least probability scores is effectively divided into 5 equal pieces to create effective fuzzy categorization labels. These designations pertain to the fuzzy crisp values, which have been classified as extremely high, medium-low, or very low.

The neurons of the Artificial Neural Network are therefore classified using the following criteria based on the TF-IDF, BoW and Entropy scores, with labels such as Very high, high, medium, low, and very low. The extremely high rating correlates to a high chance of hate speech, which is subsequently efficiently categorized in the following and final stage utilizing decision making. The algorithm 2 shown below may explain the complete Fuzzy ANN procedure.

ALGORITHM 2: Hidden Layer Estimation

```
//Input: Feature List $F_L$, Weight set $W_S$= { }
//Output: Hidden Layer value list $H_{LV}$
hiddenLayerEstimation ($F_L$, $W_S$)
Start
$H_{LV}$ =Ø {Hidden Layer value]
for i=0 to size of $F_L$
    ROW= $F_{L[i]}$
      for j=0 to size of ROW
          $X$=0
          for k=0 to N [Number of Neurons]
              ATR=ROW[j]
              $X = X$ + (ATR* $W_{S[index]}$)
              index++
          end for
        $H_{LV=}$ reLUmax(0, $X$)
   end for
```

```
    end for
    return $H_{LV}$
Stop
```

Step 6: Decision Making – The characteristics retrieved from the collected string are used to detect hate speech in the previous stage using the fuzzy Artificial Neural Network. The extremely high score values obtained from the Fuzzy Artificial Neural Networks are being used to provide excellent hate speech identification. However, these numbers are inconclusive and may produce false positives and other inconsistencies. As a result, these occurrences must be effectively categorized before being shown to the user. As a consequence, the Decision Making technique successfully categorizes the data using the If-then rules and shows the appropriate hate speech recognition output to the user via the Interactive User Interface.

IV RESULTS AND DISCUSSIONS

The proposed approach for hate speech estimation based on NLP and deep learning was implemented by developing the methodology in Java using the NetBeans IDE. The laptop utilized for the deployment had a typical setup with an Intel Core i5 CPU, 8GB of RAM, and a 1TB hard drive. To meet the storage requirements, the MySQL database was employed.

This technique has undergone rigorous testing in order to appropriately assess the suggested approach's performance. The precision and recall concept was used to evaluate the performance characteristics.

Performance Evaluation through Precision and Recall

Precision and recall are two really useful approaches for understanding how correctly a certain component in our methodology is used. A module's accuracy defines its relative correctness and offers a wide range of reliability.

The precision metric was calculated using our technique as the ratio of correct hate speech predictions to total messages received. The recall criteria, on the other hand, supplement the accuracy metric and assist in determining the exact effectiveness of the Fuzzy Artificial Neural Network component.

This procedure calculates recall as the proportion of correct hate speech estimations to total number of inaccurate hate speech estimations. The following equations quantitatively describe this process.

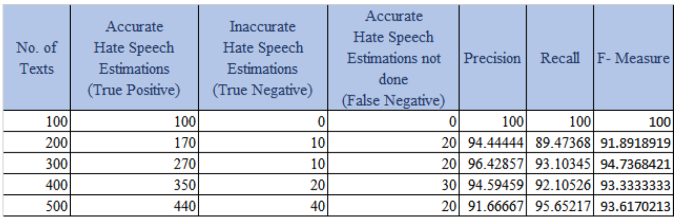
Precision can be depicted as below

- ✓ TP (True Positive) = The number of accurate hate speech estimations for the given input texts
- ✓ FP (False Positive) = The number of inaccurate hate speech estimations for the given input texts
- ✓ FN (False Negative) = The number of accurate hate speech estimations that are not done for the given input texts

So, precision can be defined as

Precision = (TP / (TP + FP)) *100
Recall = (TP / (TP + FN)) *100
F Measure = 2*(Precision*Recall)/ (Precision + Recall)

Table 1 below summarizes the empirical findings acquired using the aforementioned formula. Figure 2 shows how these tabular information are combined to produce a visual representation.



No. of Texts	Accurate Hate Speech Estimations (True Positive)	Inaccurate Hate Speech Estimations (True Negative)	Accurate Hate Speech Estimations not done (False Negative)	Precision	Recall	F- Measure
100	100	0	0	100	100	100
200	170	10	20	94.44444	89.47368	91.8918919
300	270	10	20	96.42857	93.10345	94.7368421
400	350	20	30	94.59459	92.10526	93.3333333
500	440	40	20	91.66667	95.65217	93.6170213

Table 1: Precision and Recall Measurement Table

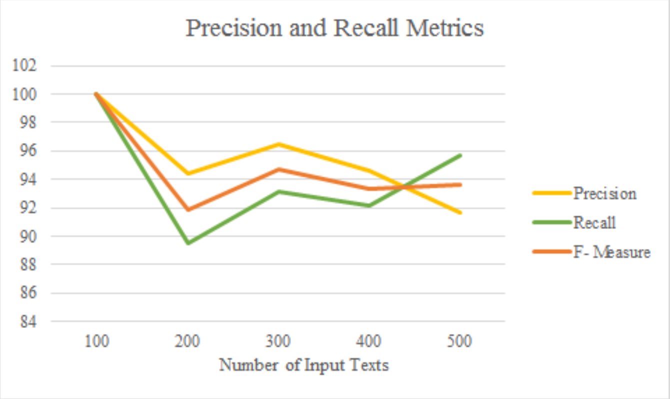


Figure 2: Comparison of Precision, Recall &F-Measure

The graph demonstrates the Fuzzy Artificial Neural Network's efficacy in predicting hate speech associated with the input texts. The approach's outstanding efficiency is demonstrated by precision and recall ratings of 95.42 percent and 94.06 percent, respectively. These figures are pretty useful and satisfying for a first-time application of such a methodology.

The precision, recall, and accuracy scores examined for hate speech detection revealed the suggested system's usefulness in great detail. The proposed method was successfully compared to the methods described in [15]. Our approach has a precision of 95.42 percent and an accuracy of 94.71 percent. The correlation of the graph-based hate speech detection strategy with the proposed methodology is shown in table 2 below in a tabular style.

Performance Metric	Our approach (Fuzzy ANN)	Graph Based approach [15]
Precision	95.42	81
F - Measure	94.71	67

Table 2: Precision, and Accuracy comparison

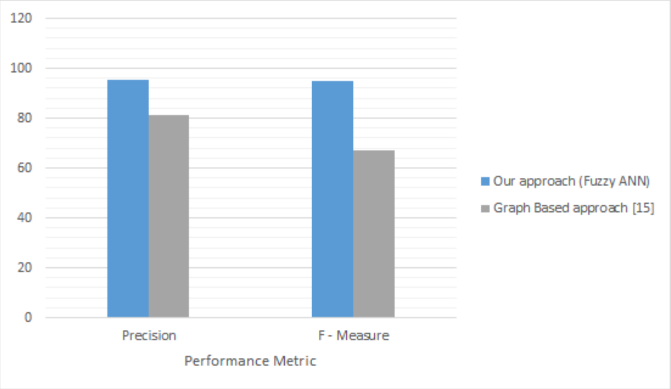


Figure 3: Comparison with Graph based technique depicted in [15]

As seen in Figure 3, the deep learning methodology proposed in this research paper outperforms the graph-based hate speech identification approach proposed in [15]. This is owing to the Fuzzy Artificial Neural Network that has been implemented to significantly improve the accuracy of hate speech identification. These findings are extremely satisfactory because the given system achieves the accuracy indicated by the performance scores.

V CONCLUSION AND FUTURE SCOPE

This research study presents an effective technique for detecting hate speech on various media platforms. This method efficiently takes the string from either the audio and adds it into a workbook which is provided as an input to the system which initially preprocesses it. Following preprocessing, the preprocessed text is delivered for Natural Language Processing utilizing the string's Term Frequency and Inverse Document Frequency. The TF-IDF and preprocessed string are then used to estimate entropy. Entropy is estimated utilizing Shannon Information Gain, which yields entropy values that are fed into Fuzzy Artificial Neural Networks for neuron synthesis. The Fuzzy ANN technique is charged with identifying hate speech, which yields likelihood ratings. To estimate hate speech, these likelihood scores are successfully categorized using the Decision Making technique. The approach's execution effectiveness has been properly assessed through a number of tests, yielding good results of precision and Recall. The proposed method yields a precision of 95.42 percent and an accuracy of 94.71 percent over the precision and accuracy of 81% and 61% of graph based approach as mentioned in [15].

In the future this model can be enhanced to detect the live hate speech in audio, video, tweets and comments by deploying the generative adversarial neural network in cloud paradigm.

REFERENCES

[1] K. A. Qureshi and M. Sabih, "Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text," in IEEE Access, vol. 9, pp. 109465-109477, 2021, DOI: 10.1109/ACCESS.2021.3101977.

[2] A. Rodriguez, Y. -L. Chen and C. Argueta, "FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis," in IEEE Access, vol. 10, pp. 22400-22419, 2022, DOI: 10.1109/ACCESS.2022.3151098.

[3] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," in IEEE Access, vol. 8, pp. 204951-204962, 2020, DOI: 10.1109/ACCESS.2020.3037073.

[4] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López and M. T. Martín-Valdivia, "A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis," in IEEE Access, vol. 9, pp. 112478-112489, 2021, DOI: 10.1109/ACCESS.2021.3103697.

[5] Ashwin Geet d'Sa, Irina Illina, and Dominique Fohr, " Classification of Hate Speech Using Deep Neural Networks" in HAL Open Access, HAL Id: hal-03101938 https://hal.archives-ouvertes.fr/hal-03101938.

[6] C. Baydogan and B. Alatas, "Metaheuristic Ant Lion and Moth Flame Optimization-Based Novel Approach for Automatic Detection of Hate Speech in Online Social

Networks," in IEEE Access, vol. 9, pp. 110047-110062, 2021, DOI: 10.1109/ACCESS.2021.3102277.

[7] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep Learning-Based Fusion Approach for Hate Speech Detection," in IEEE Access, vol. 8, pp. 128923-128929, 2020, DOI: 10.1109/ACCESS.2020.3009244.

[8] M. Mozafari, R. Farahbakhsh, and N. Crespi, "Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta-Learning," in IEEE Access, vol. 10, pp. 14880-14896, 2022, DOI: 10.1109/ACCESS.2022.3147588.

[9] M. Z. Ali, Ehsan-Ul-Haq, S. Rauf, K. Javed, and S. Hussain, "Improving Hate Speech Detection of Urdu Tweets Using Sentiment Analysis," in IEEE Access, vol. 9, pp. 84296-84305, 2021, DOI: 10.1109/ACCESS.2021.3087827.

[10] O. Oriola and E. Kotzé, "Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets," in IEEE Access, vol. 8, pp. 21496-21509, 2020, DOI: 10.1109/ACCESS.2020.2968173.

[11] H. S. Alatawi, A. M. Alhothali, and K. M. Moria, "Detecting White Supremacist Hate Speech Using Domain-Specific Word Embedding With Deep Learning and BERT," in IEEE Access, vol. 9, pp. 106363-106374, 2021, DOI: 10.1109/ACCESS.2021.3100435.

[12] L. H. Son, A. Kumar, S. R. Sangwan, A. Arora, A. Nayyar, and M. Abdel-Basset, "Sarcasm Detection Using Soft Attention-Based Bidirectional Long Short-Term Memory Model With Convolution Network," in IEEE Access, vol. 7, pp. 23319-23328, 2019, DOI: 10.1109/ACCESS.2019.2899260.

[13] V. -I. Ilie, C. -O. Truică, E. -S. Apostol and A. Paschke, "Context-Aware Misinformation Detection: A Benchmark of Deep Learning Architectures Using Word Embeddings," in IEEE Access, vol. 9, pp. 162122-162146, 2021, DOI: 10.1109/ACCESS.2021.3132502.

[14] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018, DOI: 10.1109/ACCESS.2018.2806394.

[15] M. Beatty, "Graph-Based Methods to Detect Hate Speech Diffusion on Twitter," 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2020, pp. 502-506, doi: 10.1109/ASONAM49781.2020.9381473.