NLP based Analysis and Detection of Unethical Text

Apurva Khandekar
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
khandekarapurva@gmail.com

Chekuri Devi Hema
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
devihemachekuri@gmail.com

Alle Meghana
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
meghanaalle10@gmail.com

Akuraju Mounika
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
mounikaakuraju@gmail.com

V. S. Vaishnavi
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
sonyvaishu2607@gmail.com

***Abstract*—In the present scenario, the internet has enabled everyone to express their opinion. As every coin has two sides, it paved a way for increased negativity and intolerance. In this context, unethical text includes offensive comments and hate speech targeting a group or an individual based on their characteristics such as race, religion or gender and that may threaten social peace. So, the unethical text should be eliminated before it reaches a user where identifying an immoral text plays a pivotal role. There are various methods that are used in identifying a text, one such method, widely known as Natural Language Processing (NLP) is used in this project for the text analysis in the process of detection of unethical text. NLP enables us to perform tokenization, embedding matrix, padding and identification of semantic relationships which helps us to analyse the text data efficiently. Different algorithms were used to detect the unethical text, in this project we have used LSTM and Bi-LSTM to identify whether the user's statement is offensive or not. The better and accurate model is taken and a web page is built on it. It predicts the offensiveness of text present in user input. In this way, our project detects the unethical text on the web which can be used to eliminate unfiltered comments. The best model is achieved by using the Bi-LSTM with an accuracy of 86.4.**

Keywords*— *Natural Language Processing, Bidirectional Long Short Term Memory, Deep Learning, Hate speech detection, Unethical text, social media negativity.

I. INTRODUCTION

In today's world, social media content is increasing day by day. Billions of people post content every day in the form of images, audio, video and text. Most of the social media platforms like Instagram, YouTube, Twitter allows users to express their opinions publicly. These opinions can be negative or a positive statement towards a particular individual or a group which can be in the form of comments or a counter audio or video. And negative statements can weigh a huge impact and can result in intolerance in society and mental health aggravation in people [1]. So, it is important to detect unethical texts.

Advancements in Artificial intelligence, machine learning and usage of its algorithms and structures will aid us in classifying the text with more precision. It is pivotal to classify the unethical text from ethical texts. In social media most of the text is unstructured, classifying such text is demanding because of its context-dependent interpretation of natural language. This can be handled by using data mining techniques [2]. Natural Language Processing (NLP) is a field consisting of different computational operations in text mining which aids in comprehending the human natural language by the computer or machine [3].

In NLP we have different types of techniques to detect hateful and unethical text such as Part- of-Speech (PoS), Dictionaries, Bag-of-Words (BOW), Term Frequency-Inverse Document Frequency (TF-IDF), these techniques are widely used in listing the words sequentially and determining its importance [4,5]. One of the basic techniques is tokenization. Tokenization divides a sentence or a document into words or lines in the form of tokens [6].

In existing works there are usage of algorithms such as Logistic regression, Naive Bayes (NB) and support vector machine [7]. In the starting phase the researchers tried to apply word uni-gram features to the NB algorithm where the model was only able to detect the unethical text irrespective of the context [8]. In continuation there are projects that are done on LSTM [9]. Bidirectional LSTM (Long short-term memory) is used to train the model which is almost similar to the functionality of regular LSTM except there is a difference in the input flow [10].

II. RELATED WORK

A. *Observations from previous literature work*

The classical machine learning algorithms like Logistic Regression, Naïve Bayes, Support Vector Machine and Random Forest are used in the study of hate speech detection in the existing literature. As the attached figure shows, machine learning algorithms produced comparatively low accuracies compared to the deep learning algorithms.

Using neural networks, there is continuous improvement in the performance of the models from Multi-Layer Perceptron to CNN (which is generally preferred for image data) to recent developments using LSTM and Bi-LSTM. Accuracies in below figure are shown in %.

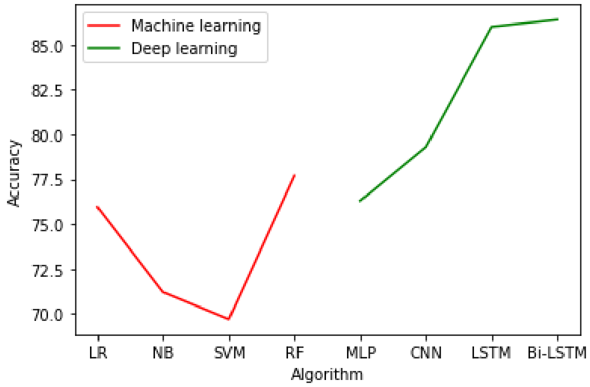

Fig. 1. Analysis with various algorithms

From the above analysis, we can conclude that the accuracy increases from machine learning algorithms to deep learning algorithms, further from CNN to LSTM to our proposed model using Bi-LSTM.

B. *Limitations of existing work*

Major drawbacks of the existing models are the memory constraints that occur with the real-time social media dataset while using machine learning algorithms which increases training time and slows down the efficiency. Considering the 'context' aspect while dealing with text data is crucial which undermines CNN, RNN and LSTM compared to Bi-LSTM.

III. METHODOLOGY

A. Architecture of the proposed system

The proposed system consists of the web application which takes the user statement as input and predicts whether it is categorized under "Unethical" or not. For this, the Bidirectional Long Short-Term Memory (Bi-LSTM) algorithm is used and as human text is analysed, we make use of Natural Language Processing (NLP) techniques like tokenizer [11]. The existing systems on this work are concerned with classical machine learning algorithms like Logistic Regression, Naïve-Bayes and we find the evolution towards deep learning with the help of CNN and LSTM in recent times. Thus, we used Bi-LSTM to overcome the shortcomings of the existing systems [10].

B. Algorithm selection and Pseudocode

From the analysis of existing literature work, we identified that classical machine learning algorithms have been widely used for hate-speech detection and related research work dealing with text based data.

But deep learning algorithms which are more efficient are rarely used which gives us the scope to produce real-time analysis. Bi-LSTM is thus selected to overcome the limitations of memory and speed constraints due to large datasets along with context-based analysis which is possible through bidirectional nature of this algorithm compared to the LSTM.

The flow of proposed system goes as follows:

- Uploading the dataset for training purpose

- Performing NLP via tokenizer, embedded matrix and padding techniques
- The packages are imported from TensorFlow and Keras. The data is trained and tested using Bi-LSTM algorithm
- In the algorithm, there is usage of Adam optimizer, Softmax activation function, categorical Cross-Entropy loss function and epochs set to 15
- Integrating backend with the frontend of the proposed system
- Taking user input statement
- Predicting the output whether the statement is unethical or not
- Producing the results.

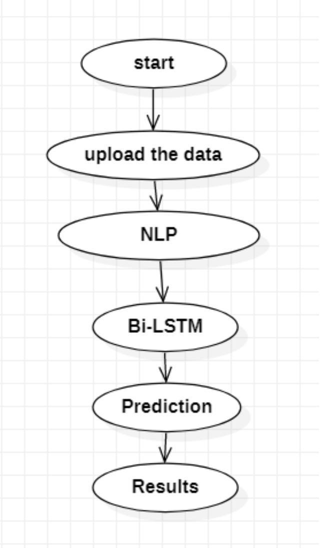

Fig. 2. Architecture of the proposed system

C. Dataset

The data collected from specific social media platform like Twitter (tweets) or YouTube (comments) is not used for this application. The data used consists of diversified user statements which are refined and available on Kaggle platform.

Data considered for training has 40000+ text statements in which ~ 60% falls under the label "Hate" and ~ 40% comes under "Not Hate". The train-test split is 75:25 and the model produced efficient results with this diversified and processed dataset.

D. Working model

Initially, the required packages are imported from TensorFlow and Keras to train the model using Bi-LSTM. As our system involves web application, we used "Flask" package to integrate the frontend and backend including model training and NLP analysis [12]. It is a simple web application with user interactivity consisting of a single web page which inputs user text and predicts if it's unethical or not.

E. Mathematical model

Current state = h(t)
Previous state = h(t-1)
Input state = x(t)
Weight at input layer = w(xh)
Weight at recurrent neuron = w(hh)
Weight at output layer = w(hy)

Input function:

$$h(t) = f\ (h(t-1), x(t))$$

Activation function:

$$h(t) = \tanh\ ((w(hh)h(t-1)) + (w(xh)(x(t)))$$

Output function:

$$y(t) = w(hy)h(t)$$

F. Pre-processing via NLP analysis

NLP or the Natural Language Processing techniques are used for the effective pre-processing as we deal with a huge dataset of texts [3]. Tokeniser, embedded matrix and padding are the steps that are performed.

- Tokenisation is done by simply breaking down the text statements into number of tokens which is an essential step in text analysis to determine how efficiently the model can interpret with the user input [6].

  Example-
  Input: 'Targeting a specific individual based on the community they belong is unacceptable.'
  Output: ['Targeting', 'a', 'specific', 'individual', 'based', 'on', 'the', 'community', 'they', 'belong', 'is', 'unacceptable']

- Embedded matrix is used to store the word in a form compatible with the machine to interpret, by attaching specific embeddings to the tokens and storing them in real-valued vector for further steps. In this way, the words which are near-by in the matrix are expected to be similar in their meanings.

- Padding is performed as the final step to standardize the length of all the sequences of data considered for training the model. It is done by adding required number of zeroes in the beginning or at the end of sequential data.

By performing the above techniques, pre-processing is achieved in our proposed model.

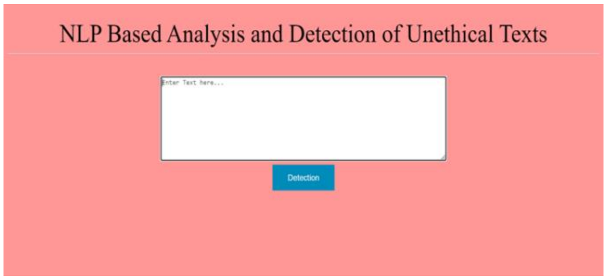

Fig. 3. Web application of proposed system

IV. RESULTS

As the training is done in the proposed model, we observed the progress in accuracy and decline in loss values with the increasing epochs. By considering the trade-off between training time and number of epochs, it has been set to 15. This can be observed from the attached image. The final accuracy achieved is 86.98% and this highly depends on the dataset considered. Unlike the existing systems which mainly depend on dataset from specific social media platforms (like Twitter tweets or YouTube comments), this application includes a diversified dataset.

TABLE 1. Numerical values of results for proposed and existing methods

	Model	Accuracy
Existing method	FastText - Wiki + CNN + GRU	0.83
Existing method	Bi-directional LSTM CNN	0.81
Existing method	Bag and sequences of words & Naive Bayes	0.76
Existing method	Characters n-grams & Support Vector Machine	0.75
Proposed method	Tokeniser, embedded matrix and padding & Bi-LSTM	0.8698

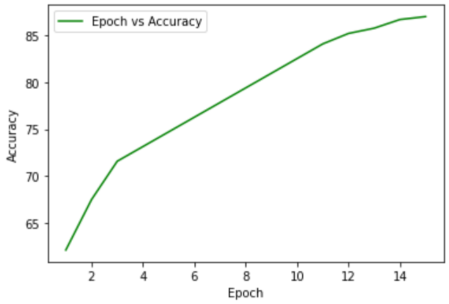

Fig. 4. Epoch vs Accuracy graph

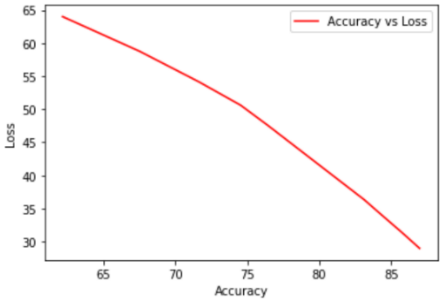

Fig. 5. Accuracy vs Loss graph

V. FUTURE SCOPE

In the proposed system the training data includes 40,000+ text statements. The performance of the system can be further improved by:

- Training the model on mixed datasets
- Increasing the number of epochs (with a balance of memory constraints)
- Resampling the model with cross-validation
- Improving the quality of training dataset.

VI. CONCLUSION

In this paper, the systematic literature review aims to identify and analyse the datasets, NLP analysis and various algorithms used in research of Hate Speech Detection from previously established works. The Proposed System detects whether the text is Unethical or not (i.e., hate or not hate). We proposed a solution for the detection of the offensive texts using deep learning Bi-LSTM. We used Bi-LSTM for the detection of hate text and NLP techniques for pre-processing the data. Our proposed approach reaches an accuracy equal to 86.98% and more by increasing the number of epochs but considering the trade-off between training time and accuracy, we fixed epochs to 15. We conclude that classical machine learning classifiers such as Naive Bayes, Logistic Regression and SVM are commonly used for supervised problem statements but deep learning classifiers such as LSTM, and

CNN seem effective with real-time data. To overcome the shortcomings of LSTM (considering the text data), Bi-LSTM is used to produce effective results.

VII. REFERENCES

[1] Blaya, C. (2018). *Cyberhate: A review and content analysis of intervention strategies. Aggression and Violent Behavior.* doi:10.1016/j.avb.2018.05.006

[2] Irfan, R., King, C. K., Grages, D., Ewen, S., Khan, S. U., Madani, S. A., … Li, H. (2015). *A survey on text mining in social networks. The Knowledge Engineering Review, 30(02), 157–170.* doi:10.1017/s0269888914000277

[3] Hirschberg, J., & Manning, C. D. (2015). *Advances in natural language processing. Science, 349(6245), 261–266.* doi:10.1126/science.aaa8685

[4] Alrehili, A. (2019). *Automatic Hate Speech Detection on Social Media: A Brief Survey. 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA).* doi:10.1109/aiccsa47632.2019.9035228

[5] Koushik, G., Rajeswari, K., & Muthusamy, S. K. (2019). *Automated Hate Speech Detection on Twitter. 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA).* doi:10.1109/iccubea47591.2019.9128428

[6] Sun, S., Luo, C., & Chen, J. (2017). *A review of natural language processing techniques for opinion mining systems. Information Fusion, 36, 10–25.* doi:10.1016/j.inffus.2016.10.004

[7] Istaiteh, O., Al-Omoush, R., & Tedmori, S. (2020). *Racist and Sexist Hate Speech Detection: Literature Review. 2020 International Conference on Intelligent Data Science Technologies and Applications (IDSTA).* doi:10.1109/idsta50958.2020.9264052

[8] I. Kwok and Y. Wang, “Locate the hate: Detecting tweets against blacks,” in Proceedings of the twenty-seventh AAAI conference on artificial intelligence, 2013, pp. 1621–1622.

[9] Rini, R., Utami, E., & Hartanto, A. D. (2020). *Systematic Literature Review Of Hate Speech Detection With Text Mining. 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS).* doi:10.1109/icoris50180.2020.9320755

[10] Istiake Sunny, M. A., Maswood, M. M. S., & Alharbi, A. G. (2020). *Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model. 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES).* doi:10.1109/niles50944.2020.9257950

[11] Bakar, J. A., Omar, K., Nasrudin, M. F., & Murah, M. Z. (2014). *Tokenizer for the Malay language using pattern matching. 2014 14th International Conference on Intelligent Systems Design and Applications.* doi:10.1109/isda.2014.7066258

[12] Yaganteeswarudu, Akkem (2020). [IEEE 2020 5th International Conference on Communication and Electronics Systems (ICCES) - COIMBATORE, India (2020.6.10-2020.6.12)] 2020 5th International Conference on Communication and Electronics Systems (ICCES) - Multi Disease Prediction Model by using Machine Learning and Flask API. , (), 1242–1246. doi:10.1109/icces48766.2020.9137896