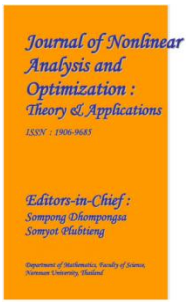
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 1, No.15 : 2024
ISSN : **1906-9685**



Paper ID: ICRTEM24_117 **ICRTEM-2024 Conference Paper**

HATE SPEECH DETECTION USING DEEP NEURAL NETWORKS FOR SAFER ONLINE ENVIRONMENT

[#1]**Dr. A. Shiva Kumar,** ***Associate Professor***
[#2]**S.Y. Ravindra Kumar,** ***UG Student***
[#3]**J. Saketh,** ***UG Student***
[#4]**P. Sai Shankar,** ***UG Student***
Department of Information Technology,
CMR COLLEGE OF ENGINEERING AND TECHNOLOGY, HYDERABAD.

Abstract: The spread of harmful content and hate speech on the internet in the modern digital age is a serious danger to the inclusion and well-being of online communities. Our project focuses on developing a website for hate speech identification using state-of-the-art artificial intelligence (AI) technology in order to address this rising issue. We go into the creation, importance, and effects of our website in this presentation. The objective of our effort is to enable individuals, social media platforms, and online communities to firmly oppose hate speech and promote a more secure online community. We will go through the main elements of our project, such as the user-friendly interface that enables real-time hate speech analysis, data gathering and categorization, and AI model training.

Key Words: *Safer online environment, Artificial Intelligence Technologies, Real-time hate speech analysis, Deep Neural Networks*.

Introduction

The prominence of social media platforms and other online forums has provided people with previously unheard-of opportunities to share their ideas, views, and feelings in the rapidly changing world of online communication. However, more connectivity across digital platforms has revealed a negative side: an increase in hate speech. Hate speech, defined as any form of communication that stigmatizes, discriminates against, or incites violence against individuals or groups based on characteristics such as race, ethnicity, religion, gender, or sexual orientation, poses a serious threat to open discourse, inclusivity, and respect.

It is critical to develop innovative and effective solutions to the challenge of recognizing and decreasing hate speech in online spaces. Our study intends to enhance current endeavors by presenting a Deep Neural Network (DNN)-based Hate Speech Detection Model. The goals of this model are to precisely identify hate speech while also developing a versatile tool that can be readily integrated into other websites to provide a robust barrier against the spread of offensive and discriminatory content.

This paper discusses the motivation for developing our hate speech detection model, the relevance of eliminating hate speech online, and the approaches utilized to create a DNN-based solution. As we navigate the complexities of hate speech detection, we look at the challenges of distinguishing between offensive language and protected speech, the nuances of context in online debates, and the moral quandaries that occur when utilizing content moderation technologies.

Research Objective: Our goal is to provide a complete understanding of our Hate Speech Detection Model, its performance indicators, and its potential contribution to the formation of a more inclusive and secure online community by the end of this study. The demand for effective tools to prevent hate speech is increasing as the digital domain continues to transform how we interact with one another in society. With this study, we seek to advance the common objective of building online settings in which a variety of perspectives can coexist peacefully, free of prejudice and hatred.

Related Work

Detecting Hate Speech and Using It to Examine Immigration-Related Events in ItalyThe paper offers a novel method for examining immigration-related phenomena that combines two sources of information: one from the analysis of naturally occurring hate speech on immigrants on Twitter using automatic hate speech detection techniques, and the other complementary from a selection of pertinent official survey-based statistical demographic data that is periodically made available by national institutes, including a set of intriguing We look at Italy, a nation in Europe, as a case study. A recent natural decline in Italy's population was completely countered by net migration, which accounted for 108% of the overall change in population [3]. We think it is particularly crucial to test our methods in countries with comparable demographics in order to better understand immigration and related issues. In particular, we are curious about the growing correlations between crime, education, and employment indices as well as the prevalence of hate speech in the community's online discourse. We present our data in this way, which parallels the North-South socioeconomic divide in Italy and suggests a connection between economic and cultural traits and hate speech online.

The ability of a system to learn without explicit programming is known as machine learning, according to the study [2] Automated Hate Speech Detection on Twitter. The system is trained using a variety of machine learning approaches, and as it acquires experience, it can automatically learn and get better.

A branch of artificial intelligence known as spoken language processing enables a computer system to comprehend spoken language that is spoken by people. But machines aren't very good at comprehending emotions. Sentiment analysis is a method for identifying a language's positive, negative, or neutral feelings. Sentiment analysis employs natural language processing to ascertain the sentence's polarity. In this study, we developed a machine learning model that can distinguish between hate speech and non-hate speech in tweets using natural language processing in Python. The bag of words and TFIDF (term frequency-inverse document frequency) features of a publicly accessible Twitter dataset were extracted and used to train the logistic regression classifier. According to our findings, the classifier has a 94.11% accuracy rate in determining if a tweet is nasty or not.

Certain features of hate speech prevent automatic identification without the help of human annotators, according to the author of [3] MC-BERT4HATE: Hate Speech Detection utilizing Multi-channel BERT for Different Languages and Translations. First off, there are no hard and fast rules about what defines hate speech. Languages that are considered objectionable vary in frequency according on the nation, period, society, and level of political development. Consequently, combining many hate speech datasets annotated with different criteria to increase data size is challenging. Second, hate speech cannot be detected by merely checking for swear words in comments. For instance, sarcasm is commonly used in hate speech; therefore,

understanding its context and subtleties is crucial. Therefore, in order for the hate speech classifier to accurately identify the purpose of the material, it must be able to identify a wide range of attributes. Despite these drawbacks, a lot of research was done to find a solution. Hate speech detection is often considered a binary classification task, while more granular categories can be incorporated, like predicting the kind of hate speech or its degree of aggression. The challenge as a supervised sentence or document categorization task was the main focus of previous research [4]. Some used logistic regression, support vector machines (SVM), and Naive Bayes classifiers using the updated features after feature engineering. Others used the deep learning paradigm, which automatically finds hidden features through the use of deep neural network architectures. However, prior research has paid little attention to approaches based on transfer learning.

[4] Twitter Dataset for Indonesian Language Cyberbullying and Hate Speech Detection.

An report on kata.co.id claims that Indonesia is seeing a 50% increase in internet users in only a single year. Indonesia has surpassed the global growth rate of 10% to become the country with the second-highest internet user base in the world. In January 2018, a study by We Are Social found that 132.7 million people in Indonesia utilize the internet out of the country's 265.4 million total population. 130 million people are known to be frequent users of social media out of all internet users. It may be inferred from these results that social media is used by almost all internet users. Another technologically summarized fact. Techno.okezone.com reports that Indonesians use various devices to access the internet for an average of 8 hours and 51 minutes per day. The typical user uses several devices to spend 3 hours and 23 seconds on social media out of this total. These data show that among Indonesian citizens, social media is one of the most popular content kinds. Internet users can now access social media whenever and whenever they want, with an average fixed broadband connection speed of 13.78 Mbps and mobile broadband rates of 9.82 Mbps.

[5] Using the maximum entropy classification method, hate speech in Indonesian language is detected in the Instagram comment section. Modern technology has given rise to a new platform via which we can express our opinions. Inciting violence or hatred against groups based on particular characteristics, such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity, or others, is considered hate speech, according to Paula Fortuna's 2017 thesis on the subject. Hate speech can take many different forms, including subtle forms or humor[1]. The definition of hate speech in Indonesia is different. According to Section 28 subsection 2 of the Indonesian ITE constitution, "people who spread hate speech are every person who deliberately and without right spread information that causes hatred among an individual/group of people based on ethnic, race, religion, and class." This statement applies to hate speech on social media. Many people have come together in support of this constitution since hate speech on social media has increased. Because of their range and regularity, hate speech is one of the concerns that authorities find difficult to deal with. On social media, hate speech can appear in a variety of formats, such as videos, images, and comments. The number of hate speech prosecutions in Indonesia has rapidly increased, especially after the 2019 presidential election and the 2018 election of the Jakarta governor. According to Minister of Technology, Communication, and Informatics Rudiantara, radicalism, terrorism, and other forms of hate speech may now be found on social media sites like Facebook and Instagram.

Within [6] SEMAR: A Machine Learning-Based Interface for Identifying Hate Speech in IndonesiaAn automated system for detecting hate speech in Indonesia is proposed by this work, called SEMAR. The Javanese puppet character Semar served as the inspiration for this name. Semar was chosen as the name because he is a figure who exercises extraordinary caution in all aspects of his behavior, including speech and mental expression. In addition to its automatic hate comment detection capabilities, SEMAR is designed to guarantee that training data keeps evolving. Users can remark on the engine's predictions using this technique, and their comments are stored in the database as fresh training data. Every day, the System will carry out self-training using both historical and new training data, enabling the model's performance to advance with time. The purpose of this project is to develop an Application Programming Interface (API) for

machine learning classification models in order to recognize hate speech. The API will then be utilized to develop a Word Press plugin that blocks hate comments.

Within [7] As per the author of HateSense: Tackling Ambiguity in Hate Speech Detection, the unchecked spread of hate speech on the internet and the dearth of effective methods to censor such expressions have caused numerous negative effects on society and are broad issues. Numerous investigations are underway about the automated identification of hate speech on social media networks. In order to solve the issue of hate speech, researchers treated this as any other text mining problem, utilizing machine learning models and natural language processing techniques. Whether applied singly or in combination, these strategies have shown to be successful in identifying hate speech on the internet. But hate speech is a multifaceted, highly subjective problem with a wide range of linguistic nuances. Because of this, existing systems that employ general text mining algorithms have not been able to resolve the ambiguity that frequently arises and have not been able to accurately determine the context in which the hate speech was expressed.

Proposed System

Our research's main goal is to identify and filter hate speech on social media sites, as it is mentioned in the abstract.The primary challenge with existing technologies is their restricted use and availability.Of course, our solution is just another Hate Speech Detection model, but we want to set ourselves apart from the competition by letting other clients incorporate it into their applications.TensorFlow's capabilities are being used to train our model to detect hatefulness using the dataset. The words will then be divided into a bag of words using natural language processing (NLP).The DNN algorithm will now be used to determine if a communication constitutes hate speech.

Methodologies:

Data Collection: Provide a full overview of the platforms, forums, or websites where your training data was collected.

Text Representation:

Describe the methods for converting unprocessed text data into a format that may be used to train a neural network, such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words (BoW).

Neural Network Architecture**:**

Provide an overview of the deep neural network's architecture, including the number of layers, activation function selection (such as Sigmoid or ReLU), and the number of neurons per layer.

Training Procedure:

Describe the training approach in full, including weight settings, optimization techniques (such as Adam Optimizer and Gradient Descent), and regularization strategies (such as L2 Regularization).\

Evaluation Metrics:

Describe the measurements used to evaluate the model's performance (precision, recall, F1-score, accuracy).

Integration into Websites:

Describe the approaches utilized to make the Hate Speech Detection Model adaptable and easy to integrate onto many websites.

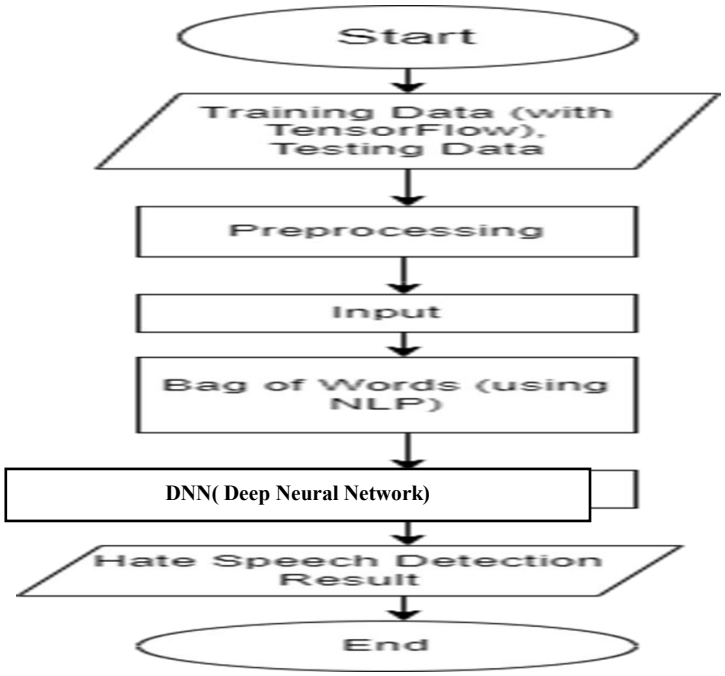


Fig 1: Block Diagram of the proposed system

Mathematical formulas for the system:

Text Representation:

1. **Bag-of-Words (BoW):***xi*=Count(wi,document)*xi* =Count(*wi* ,document)
2. **Term Frequency-Inverse Document Frequency (TF-IDF):** TF-IDF(*wi*,document)=TF(*wi*,document)×IDF(*wi*)TF-IDF(*wi* ,document)
   Where,
   xi = *i*th element in the sequence
   wi= weight of the element in model
   TF= Term Frequency
   IDF = Inverse Document Frequency

System Requirements

1. Programming Language (python 3)
2. Deep Learning Framework (TensorFlow)
3. CPU (Quadcore processor)
4. GPU (optional for improved performance)

Results & Discussions

Here, we present the findings from the tests done to evaluate our Hate Speech Detection model's effectiveness. We use the criteria provided in the approach to assess the model's capacity to discriminate between instances of hate speech and non-hate speech. With an accuracy rate of 84—already higher than most systems—we accomplished this and are currently trying to make it even better.

.

Fig 2: Result of proposed system

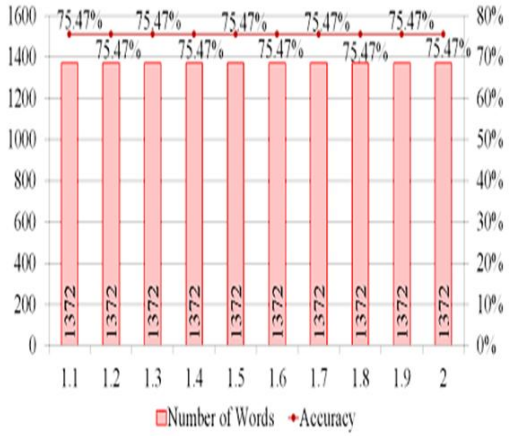


Fig 3: Classification accuracy (X) and number of words collected (Y) From [1]

Fig 4: Evaluation metrics (X) and their values (Y) for proposed system

Conclusion

In this research project, we aimed to address the pervasive problem of hate speech in online communication by developing a robust Deep Neural Network (DNN)-based Hate Speech Detection Model. The growth of hate speech on the internet poses a severe threat to the values of respect, inclusion, and free speech, so developing practical and expandable mechanisms for content moderation is critical.

In conclusion, our Hate Speech Detection Model is a step forward in leveraging cutting-edge technologies to limit the negative impacts of hate speech in virtual environments. As we move forward, we hope that this research contributes to the ongoing discussion about the ethical application of AI and lays the framework for future advances in establishing a digital world that values variety, comprehension, and peaceful debate.

References

[1] Florio, K., Basile, V., Lai, M., & Patti, V. (2020, September). Leveraging hate speech detection to investigate immigration-related phenomena in Italy. In *2020 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 1-7). IEEE.

[2] Koushik, G., Rajeswari, K., & Muthusamy, S. K. (2020, September). Automated hate speech detection on Twitter. In *2020 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)* (pp. 1-4). IEEE.

[3] Sohn, H., & Lee, H. (2021, November). Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2021 International Conference on Data Mining Workshops (ICDMW)* (pp. 551-559). IEEE.

[4] Febriana, T., & Budiarto, A. (2021, August). Twitter dataset for hate speech and cyberbullying detection in Indonesian language. In *2021 International Conference on Information Management and Technology (ICIMTech)* (Vol. 1, pp. 379-382). IEEE.

[5] Rohmawati, U. A. N., Sihwi, S. W., & Cahyani, D. E. (2020, November). SEMAR: An interface for Indonesian hate speech detection using machine learning. In *2020 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 646-651). IEEE.

[6] Kumaresan, K., &Vidanage, K. (2021, October). Hatesense: Tackling ambiguity in hate speech detection. In *2021 National Information Technology Conference (NITC)* (pp. 20-26). IEEE.

[7] Umu Amanah Nur Rohmawati Informatics Department FMIPA, Sebelas Maret Kumaresan, K. (2020). *HateSense: Tackling ambiguity in hate speech detection–An ontology, sentiment analysis and fuzzy logic-based approach* (Doctoral dissertation).

[8] Rodríguez, S. E., Allende-Cid, H., & Allende, H. (2021). Detecting hate speech in cross-lingual and multi-lingual settings using language agnostic representations. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 25th Iberoamerican Congress, CIARP 2021, Porto, Portugal, May 10–13, 2021, Revised Selected Papers 25* (pp. 77-87). Springer International Publishing.

[9] Bilal, M., Khan, A., Jan, S., & Musa, S. (2022). Context-aware deep learning model for detection of roman urdu hate speech on social media platform. *IEEE Access*, *10*, 121133-121151.

[10] Plaza-Del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, *9*, 112478-112489.