Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, No. 1, June : 2023

INAPPROPRIATE LANGUAGE AND HATE SPEECH RECOGNITION

Mrs. S.Suganya, Assistant Professor, Dept. Of Computer Science, SSM Institute of Engineering and Technology, Dindigul.
G.Y.Asmetaa, Student 1, Dept. Of CSE, SSM Institute of Engineering and Technology.
S.Harini, Student 2, Dept. Of CSE, SSM Institute of Engineering and Technology.
S.Jeyashree, Student 3, Dept. Of CSE, SSM Institute of Engineering and Technology.
J.Sriram, Student 4, Dept. Of CSE, SSM Institute of Engineering and Technology.

Abstract
The problem of online abuse, harassment, and discrimination is a growing concern in today's digital world. To address this issue, the project "Identify inappropriate language and hate speech" aims to develop algorithms and models that can automatically detect and flag instances of inappropriate language and hate speech in text-based content. The project uses natural language processing (NLP) and machine learning techniques to analyze text and identify patterns that are associated with inappropriate language and hate speech. Large datasets of annotated text are used to train and test the models, which are designed to be applied to a wide range of online content, including social media posts, comments, and messages. The ultimate goal of the project is to create tools and technologies that can help to reduce the prevalence of online abuse and promote a more respectful and inclusive online community.

Keywords: Artificial intelligence, Natural Language Processing, Text Processing, Deep Learning, API's.

I. Introduction

The rise of social media has given everyone a platform to express their opinions and thoughts. While this has created a space for free speech, it has also led to an increase in hateful and offensive language. The internet has become a breeding ground for trolls, bullies, and hate speech. This type of content not only harms individuals but can also have a negative impact on society as a whole. The current methods used to moderate online content are not fool proof and require a lot of manual work. This leads to delays in identifying harmful content and removing it from the platform. Hence, there is a need for an automated system that can identify and flag hateful and offensive language in real-time.
The Hate Speech Recognition mini-project aims to develop a model that can detect offensive language and hate speech in social media text data. The model is trained on a dataset of Twitter data, where each tweet is labeled as either hate speech, offensive language, or non-offensive text.The project utilizes a Twitter dataset containing tweets along with their corresponding labels. The dataset is loaded using the pandas library and preprocessed to clean the text. The preprocessing steps include converting text to lowercase, removing URLs and hashtags, tokenizing the text, removing stopwords and punctuation, and lemmatizing the tokens.The preprocessed text data is transformed into numerical features using the CountVectorizer, which creates a matrix of token counts. The dataset is split into training and testing sets using the train_test_split function. A Decision Tree Classifier is then trained on the training data. The trained model is evaluated on the testing data to measure its performance in detecting hate speech and offensive language.

II. Literature

Guanyi Mou and Kyumin Lee proposed An Effective, Robust and Fairness-aware Hate Speech Detection Framework with the widespread online social networks, hate speeches are spreading faster and causing more damage than ever before. Existing hate speech detection methods have limitations

in several aspects, such as handling data insufficiency, estimating model uncertainty, improving robustness against malicious attacks, and handling unintended bias (i.e., fairness). There is an urgent need for accurate, robust, and fair hate speech classification in online social networks. To bridge the gap, we design a data-augmented, fairness addressed, and uncertainty estimated novel framework. As parts of the framework, we propose Bidirectional Quaternion-Quasi-LSTM layers to balance effectiveness and efficiency. To build a generalized model, we combine five datasets collected from three platforms. Experiment results show that our model outperforms eight state-of-the-art methods under both no attack scenario and various attack scenarios, indicating the effectiveness and robustness of our model. We share our code along with combined dataset for better future research in the year of 2021.

2.1 Impact of Deep Learning Models On Hate Speech Detection

T.Akhilesh Naidu and Shailender Kumar made a research on Impact of Deep Learning Models On Hate Speech Detection that internet one of the mediums of connectivity that is available at the doorstep, with access to the internet one gets access to many web-based platforms. An increase in the use of these platforms gives us some benefits as well as some drawbacks. One of such drawbacks is hate speech. Hate speech is a topic of concern for social media platforms. With dynamically increasing datasets manual intervention of posts is quite impossible or will be time-consuming. Hate speech detection is an automated task to detect hate speech from the input. In this paper, we have compared some deep learning models like Convolution Neural Network (CNN), Recurrence Neural Network (RNN), Long Gated Recurrent Unit (GRU), and Long-Short Term Memory. The datasets used here are publicly available. The result of our analysis shows us that GRU performed better than other basic deep learning models. The model achieved an accuracy of 92.60% with an F1 score of 81.84% for dataset (D1) and the respective values for dataset (D2) are 96.15% and 83.06% in the year of 2021.

2.2 Recognition of Hate or Offensive Tweets in the Online Communities

K.Machova, D.Suchanic and V.Maslej-Kresnakova made a study on Recognition of Hate or Offensive Tweets in the Online Communities that the paper focuses on classification of text into categories as hate speech or offensive language which represent unhealthy phenomena complicating learning and communication in online space. This classification was achieved by training a model using a deep neural network. The network was tested with different amounts of neurons in the hidden layer, with three distinctive optimizers and with various learning rates in 2020.

2.3 Abusive Language Detection in Online User Content

C.Nobata, J.Tetreault, A.Thomas, Y.Mehdad, Y.Chang reviewed that Abusive Language Detection of abusive language in user generated online content has become an issue of increasing importance in recent years. Most current commercial methods make use of blacklists and regular expressions, however these measures fall short when contending with more subtle, less ham-fisted examples of hate speech. In this work, we develop a machine learning based method to detect hate speech on online user comments from two domains which outperforms a state-of-the-art deep learning approach. We also develop a corpus of user comments annotated for abusive language, the first of its kind. Finally, we use our detection tool to analyse abusive language over time and in different settings to further enhance our knowledge of this behaviour published in 2016.

2.4 Detecting hate speech on the world wide web

Detecting hate speech on the world wide web research made by W.Warner and J.Hirschberg that approach to detecting hate speech in online text, where hate speech is defined as abusive speech targeting specific group characteristics, such as ethnic origin, religion, gender, or sexual orientation.

While hate speech against any group may exhibit some common characteristics, we have observed that hatred against each different group is typically characterized by the use of a small set of high frequency stereotypical words; however, such words may be used in either a positive or a negative sense, making our task similar to that of words sense disambiguation. In this paper we describe our definition of hate speech, the collection and annotation of our hate speech corpus, and a mechanism for detecting some commonly used methods of evading common "dirty word" filters. We describe pilot classification experiments in which we classify anti-semitic speech reaching an accuracy 94%, precision of 68% and recall at 60%, for an F1 measure of. 6375 at the year of 2012.

III. Conclusion

In conclusion, this model aimed to develop a system for identifying offensive and hate speech in social media using machine learning techniques. The proposed system utilizes natural language processing techniques for preprocessing the text data and feature extraction. The decision tree classifier algorithm was employed for training the model and making predictions. The results obtained from the evaluation of the system on the test data showed promising accuracy and precision levels. The system was able to correctly identify offensive and hate speech in social media with a high degree of accuracy. This could prove to be useful for social media companies to identify and remove such content from their platforms. Extend the model to handle multiple languages and improve its performance in detecting hate speech across different languages. This could involve training the model on multilingual datasets and incorporating language-specific features. Enhance the model to consider the contextual information of the text. Context plays a crucial role in determining whether a particular statement is hate speech or not. Incorporating contextual analysis techniques such as sentiment analysis, entity recognition, or topic modeling can provide a more accurate assessment of hate speech. Currently, the model classifies text into binary categories (hate speech or not). Consider expanding the classification to include finer categories such as different types of hate speech or offensive language. This can provide more detailed insights and help in better understanding the nature and nuances of offensive content.

References

[1] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad and Y. Chang, "Abusive language detection in online user content", WWW, 2016, [online] Available: https://doi.org/10.1145/2872427.2883062.
[2] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic and N. Bhamidipati, "Hate speech detection with comment embeddings", WWW, 2015, [online] Available: https://doi.org/10.1145/2740908.2742760.
[3] P. Badjatiya, M. Gupta and V. Varma, "Stereotypical bias removal for hate speech detection task using knowledge-based generalizations", WWW, 2019, [online] Available: https://doi.org/10.1145/3308558.3313504.
[4] Z. Zhang, D. Robinson and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network", European semantic web conference, 2018, [online] Available: https://doi.org/10.1007/978-3-319-93417-4_48.
[5] T. Davidson, D. Warmsley, M. Macy and I. Weber. "Automated hate speech detection and the problem of offensive language", Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, 2017.
[6] W. Warner and J. Hirschberg, "Detecting hate speech on the world wide web", Second workshop on language in social media, 2012.
[7] A. Arango, J. Pérez and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation", SIGIR, pp. 45-54, 2019, [online] Available: https://doi.org/10.1016/j.is.2020.101584.
[8] I. Kwok and Y. Wang, "Locate the hate: Detecting tweets against blacks", Proc. 27th AAAI Conf. Artif. Intell., pp. 1621-1622, 2013.
[9] Hao Chen, Susan McKeever and Sarah Jane Delany, Abusive Text Detection Using Neural Networks., AICS, 2017.

[10] G. K. Pitsilis, H. Ramampiaro and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks", Appl. Intell., vol. 48, pp. 4730-4742, 2018.
[11] Prashant Kapil and Asif Ekbal, "A deep neural network based multi-task learning approach to hate speech detection", Knowledge-Based Systems, pp. 106458, 2020.
[12] B. R. Amrutha and K. R. Bindu, "Detecting Hate Speech in Tweets Using Different Deep Neural Network Architectures", 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 923-926, 2019.
[13] S. Zimmerman, C. Fox and U. Kruschwitz, "Improving Hate Speech Detection with Deep Learning Ensembles", Proceedings of the LREC 2018 - Language Resources and Evaluation Conference 2018, pp. 2546-2553, 2018.
[14] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, "Bag of Tricks for Efficient Text Classification", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, pp. 427-431, April 2017.