ISSN NO: 2249-3034

OFFENSIVE LANGUAGE DETECTION IN TWEETS

Dr.THOMAS FELDMAN

[125]Assistant Professor, [346]Associate Professor

[1246]Department of Artificial Intelligence & Data Science, [3]Department of Artificial Intelligence &Machine Learning, [5]Department of Computer Science Engineering,[123456]Ramachandra College of Engineering

ABSTRACT:

This article aims to use ML classification techniques to identify tweets that include inappropriate language. We compare the results of several well-known classification algorithms using a training and prediction pipeline to find the one that works best. To train our classifiers and regression models, we will use datasets collected from Twitter annotations related to hate speech and offensive language detection. using matplotlib to visualize the results of our evaluation on a publicly available 25K tweets dataset, we tuned the optimal algorithm by considering performance and time complexity in terms of metrics like accuracy, precision, and recall in both the training and test sets of data.

INTRODUCTION:

Expanding these conditions may mislead many people in the general public, and there are a lot of sites that recall the unpleasant terms for their material as well. In these cases, we see it as a model-like arrangement that highlights the existence of objectionable words in the provided text. Thus, our project titled "offensive language location in tweets utilising different relapse and classifier calculations" will proceed accordingly. We also used data from Twitter explanations for hate speech, offensive language locations, and other sources to train our model; these datasets are inputs to several classifiers and relapse models. We used matplotlib to display the results of an adequate appropriation of a publicly available dataset of 25K tweets, and we tuned the optimal calculation by considering execution complexity and time complexity, as well as measures like exactness, accuracy, and review in both the test and preparation data. Additionally, the model is connected to the user interface in such a way that it displays the status consequence of the objectionable text when the client is provided with it for checking. whatever the level of offensiveness may be.

LITERATURE SURVEY:

1. Our project is based on an IEEE paper published by Gabriel Araujo De Souza of Federal University and Da Costa-Abreu of Sheffield Hallam University, which discusses the use of machine learning and feature selection of metadata to detect offensive language from Twitter data.In this research, we employ ML methods such as SVM and Naive Bayes to classify tweets. And then you may say that Naive Bayes is better at prediction than SVM after you've tried various methods for attribute selection.
2. We use information from two sources

ISSN NO: 2249-3034

for our project:Speech expressing hatred Appendices for Twitter Writers: Zeerak Waseem and Dirk Hovy The collection includes around 17,000 Tweet IDs that have been tagged with sexism and racism. To get at the real tweets, we used a Twitter API query and this dataset that we obtained. The deletion or deactivation of the account caused the retrieval of about 5,900 tweets to fail. A system that can identify offensive language and hate speech Writers: Zeerak Waseem and Dirk Hovy Crowdsourcing has annotated around 25,000 tweets in the dataset. There are three categories for Tweets based on the amount of people who have tagged them: hate speech, objectionable language, and neither. The dataset was retrieved from GitHub as a.csv file using Python. Approach Under Consideration: In this study, we introduce the suggested model and show how it may overcome all the problems with the existing system. We trained a model to identify the client's abusive tweets in a selected dataset using SGD classifiers and then tested it using a dataset that had been pre-processed with a higher degree of accuracy. Regardless of how insulting it may be, we used to contribute a statement from the front page to the model, and the model would then declare the yield level based on that sentence.

PREVIOUS APPROACHES:

This paper presents the current framework for offensive language discovery from Twitter data. It uses historical tweets as training data and applies ML calculations, such as SVM innocent base calculations, to the training data and model. To test the model, they use test data set preprocessing and employ various strategies for characteristic choice and accuracy. Additionally, compared to the SVM computations, IT leads to higher levels of review and accuracy in the new base.

HARDWARE COMPONENTS:

Here are the hardware requirements for building the application:
System: Pentium 42.4GHZ
Processor: core i3
Monitor:15 VGA color
RAM:4GB

SOFTWARE COMPONENTS:

Here are the software requirements for building the application:

- ✓ Python
- ✓ Flask
- ✓ Werkzeug

OS Supported:

- ✓ Windows7
- ✓ Windows XP
- ✓ Windows 8

Technologies and Languages used to Develop:

- ✓ Python

WORKING:

There are many phases involved in identifying abusive language in tweets:
The first step is to collect data. Collect all the tweets that have been classified as offensive or not offensive.
Getting ready: Remove any usernames, URLs, or special characters from the text data. Make the text more readable by changing all the capital letters to

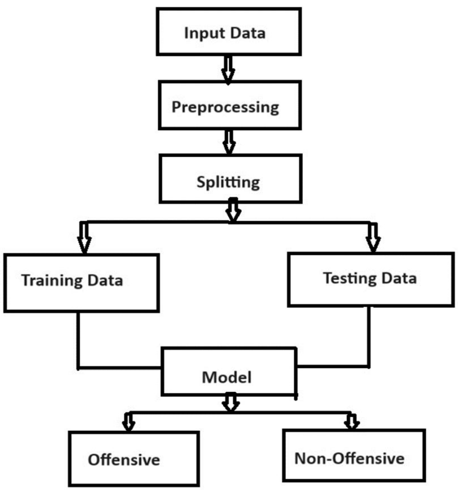
lowercase.

"Extraction of Features" Use text characteristics like bag-of-words, TF-IDF, word embeddings, and character n-grams to extract information. These characteristics provide a representation of the tweet that is amenable to methods used in machine learning.

To categorize tweets as offensive or non-offensive using the retrieved attributes, use a machine learning approach like logistic regression, support vector machines, or neural networks. Separate the dataset into two parts: training and testing. Apply the selected model to the training set of data.

Assess the model's efficacy by comparing it to the testing data using measures like accuracy, precision, recall, and F1-score. To get better results, you may need to tweak the model's hyperparameters.

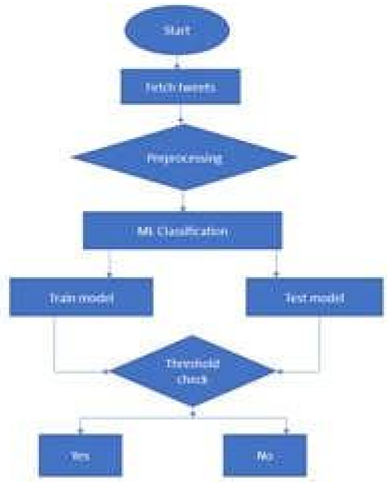
ARCHITECTURE:



IMPLEMENTATION:

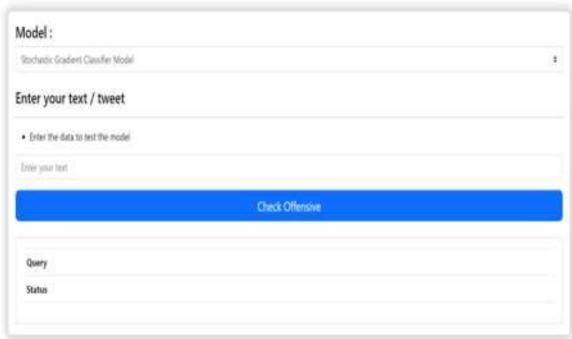
DATASET

We will train a number of classifiers and regression models using data collected from Twitter annotations related to hate speech and offensive language identification. using matplotlib to visualize the results of our evaluation on a publicly available 25K tweets dataset, we tuned the optimal algorithm by considering performance and time complexity in terms of metrics like accuracy, precision, and recall in both the training and test sets of data. Please see the Activity Diagram attached.

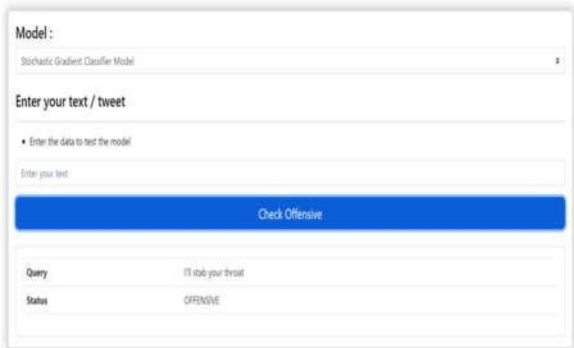


OUTPUT SCREENSHOTS:

HOMEPAGE VIEW



OFFENSIVE



ISSN NO: 2249-3034

NON-OFFENSIVE

CONCLUSION:

- In order to contribute to the problem of offensive language detection on social media platforms, we conducted a thorough literature review of previous work in the area of Offensive language detection. This allowed us to explore the novel benchmark dataset, and the following are the main conclusions that were determined.

  We developed a model to evaluate using the training dataset and investigated several approaches to dealing with the imbalance in the training set, which was used to identify objectionable tweets from a certain person in a chosen dataset. We can learn what proportion of tweets in the collection include abusive language and what proportion do not. ● By utilizing attributes such as Training Time and Prediction Time, we can categorize the Time Complexity of Algorithms. This classification can then be used to visually represent the results in terms of both algorithms and time in seconds. ● Consequently, we can learn which algorithm is most suited for the model in terms of both accuracy and time complexity. ● Additionally, we can use this information to determine if a given tweet or sentence is offensive or not.

-

REFERENCES:

[1] F. Del-Vigna, A. Cimino, F. Dell-Orletta, M. Petrocchi, and M. Tesconi, “Hate me, hate me not: Hate speech detection on facebook,” in First Italian Conference on Cybersecurity, 2017.

[2] J. Jacobs and K. Potter, Hate crimes: Criminal law & identity politics. Oxford University Press on Demand, 1998.

[3] M. Bouazizi and T. Ohtsuki, “Multi-class sentiment analysis on twitter: Classification performance and challenges,” Big Data Mining and Analytics, vol. 2, no. 3, pp. 181–194, Sep. 2019.

[4] G. Jalaja and C. Kavitha, Sentiment Analysis for Text Extracted from Twitter. Singapore: Springer Singapore, 2019, pp. 693–700.

[5] S. Sharma and A. Jain, “Cyber social media analytics and issues: A pragmatic approach for twitter sentiment analysis,” in Advances in Computer Communication and Computational Sciences, S. K. Bhatia, S.Tiwari, K. K. Mishra, and M. C. Trivedi, Eds. Singapore: Springer Singapore, 2019, pp. 473–484. [

6] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “Identifying and categorising offensive language in social media (offenseval),” arXiv preprint arXiv:1903.08983, 2019.

ISSN NO: 2249-3034

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[8] P. Liu, W. Li, and L. Zou, “Transfer learning for offensive language detection using bidirectional transformers,” in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 87–91.

[9] J. Han, S. Wu, and X. Liu, “Identifying and categorising offensive language in social media,” in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 652–656.

[10] A. Nikolov and V. Radichev, “Offensive tweet classification with bert and ensembles,” in Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 691–695.