OPEN ACCESS

Volume: 11

Special Issue: 1

Month: July

Year: 2023

E-ISSN: 2582-0397

P-ISSN: 2321-788X

Impact Factor: 3.025

Received: 08.05.2023

Accepted:13.06.2023

Published: 01.07.2023

Citation:
Deepa, KR, and S. Anusha. "Detection of Harmful Talk Utilising Machine Learning Techniques." *Shanlax International Journal of Arts, Science and Humanities*, vol. 11, no. S1, 2023, pp. 14–19.

DOI:
https://doi.org/10.34293/sijash.v11iS1-July.6309

Detection of Harmful Talk Utilising Machine Learning Techniques

Deepa K R
Department of Master of Computer Applications
Rajarajeswari College of Engineering

Anusha S
Department of Master of Computer Applications
Rajarajeswari College of Engineering

Abstract

There has been a tremendous increase in the spread of harmful content speech in today's era of internet social media platforms. They provide several enhancements. However, people with significant differences in their points of view have led to a rise in the lethality of people in internet posts and discussions. Since the pandemic's spread, companies, educational institutions, students, and the general public have all expanded their use of websites. For a long time, the increasing popularity of online platforms such as Twitter and Facebook has been a source of concern. These systems not only allow for better communication, but also allow users to share their views, which are immediately shared with the remainder of the globe. Furthermore, given the variety of these platform users' backgrounds, beliefs, ethnicity, and cultures, many of them opt to utilise derogatory, abusive, and threatening language. including classic Machine Learning and ensemble approaches. We employ a corpus acquired from the internet platform.

Keywords: Machine Learning, Toxic Speech, KNN, Linear Regression

Introduction

The internet and social networking sites are now fundamental components of how people disseminate and receive information. Social media has evolved significantly. over time, and now roughly half of people use it to communicate their ideas and thoughts. The way that individuals communicate has changed substantially over the past ten years, partly as a consequence of social media's pervasive rise. It has made it possible for a world that is more connected and informed, but it has also given way to a brand-new phenomenon: poisonous speech. Due to the availability of an open platform for the creation, discussion, and sharing of content, some quite opportunistic people have taken part in poisonous speech and generally unfavourable comments.

This design was the inspiration for our project. We want to construct a high accuracy classifier on hazardous speech using the Random Forest Classifier method to be able to efficiently identify the existence of toxic speech in comments and texts.

Because the terms "Obscene," "Toxic," "Severe Toxic," "Threat," "Insult," and "Identity Hate" are commonly used together, they have been classified as "Toxic" speech content. As a result of this, it is critical to recognise and naturally eliminate hazardous speech from internet-based social media networks. As a consequence, we developed a model with higher precision, recall, and accuracy scores for categorising given remarks into various categories of toxicity.

Literature Survey:

The distinction between hate speech and other types of speech objectionable language is a significant difficulty to detect hate speech on social media automatically. Lexical detection approaches have a poor accuracy since they identify all communications containing certain phrases as hate speech, and earlier work employing supervised learning failure to differentiate in comparison to the two groups. We gathered tweets that include hate speech terms using crowdsourced hate speech lexicon.

This study aims to investigate the many forms of venomous speech that arise on the internet by detecting profane phrases used in hate speech. This study also examines profane terms used throughout generations to help determine the user's profile. Five hundred (500) YouTube comments on hostile themes were collected. Hate speech is categorised into eight groups based on profanity.

According to the findings, 35% of the profane terms detected in our sample are connected to sexual orientation. A comparison of keywords from 1970 to 2017 reveals that sexual orientation is associated with a significant percentage of profane words. Though the present study's results are based on just 500 comments collected via a YouTube link, they are valuable in building the list.

This study aims to investigate the many forms of hatred speech that arise on internet by detecting profane phrases used in hate speech. This study also examines profane terms used throughout generations to help determine the user's profile. Five hundred (500) YouTube comments on hostile themes were collected. Hate speech is categorised into eight groups based on profanity.

This survey organises and summarises the present status of the area by offering a comprehensive summary of prior attempts, covering basic algorithms, methodologies, and key characteristics. This paper also analyses the complexities of the idea of hate speech, as defined across several platforms and settings, and offers a unified definition.

Toxic online material has become a serious concern in today's world as the usage of the internet by individuals of all ethnicities and educational backgrounds has increased exponentially. Separating hate speech from offensive language is a significant difficulty in computerised identification of hazardous text content. In this research, we present a method for automatically categorising tweets on Twitter into 3 groups: hateful, offensive, and clean.

While Websites for Social Networking facilitate communication and information exchange, they are sometimes used to start damaging campaigns against certain organisations and people. Cyberbullying, encouragement to self-harm, and Sexual predators exist, only a few of the serious consequences of enormous internet offensives. Furthermore, assaults against victim groups might develop into physical violence.

Furthermore, assaults in opposition to groups of victims might develop into physical violence. The goal of this endeavour is to restrict and prevent the worrisome spread of such xenophobic campaigns. Using Facebook as a tool, we examine the linguistic comments' content posted on a collection of the general Italian public sites. To differentiate the type of hatred, we first suggest a number of hate types.

We design and implement two classifiers for the Italian language using morpho-syntactical features, sentiment polarity, and word embedding lexicons, each based on a different learning algorithm: the first on Support Vector Machines (SVM) and the second on a specific Recurrent Neural Network called Long Short-Term Memory (LSTM).

We put these two learning algorithms to the test in order to validate their classification performance on the hate speech identification assignment. The findings demonstrate the efficacy of the two classification methods evaluated on the first manually annotated Italian Hate Speech Corpus of social media material.

Existing Model

Waseem used a 16K tweet dataset to classify them as gender stereotypes, racial prejudice, or none of the above. When compared to other approaches like as character and word n-grams, he performed the best usage of the LR algorithm. Gaydhani et al performed logistic regression using a combination of three datasets. She observed that utilising logistic regression, a term frequency and inverse document frequency vectorizer, and a term frequency and inverse document frequency vectorizer resulted in a 95.6 percent accuracy.

Davidson examined a 24k tweet corpus that he classified into three categories: hatred, offensive, and neither. He then used NLP methods to the tweets, such as extracting the base form of the phrase, text clustering, term frequency vectorization, and Darth Vader emotive procedures, before running several supervised learning algorithms, the best of which was represented LR with L-2 regularisation. They discovered, however, that using linguistic techniques, it was difficult to distinguish between Hate and Offensive content.

The precision of the current system is determined on the quality of the data.

The prediction step may be sluggish while working with large volumes of data. The current approach is sensitive to data size and irrelevant aspects.

The current approach needs a a big quantity of memory to be able to store all of the instruction data.

Proposed Methodology

Toxic commentaries are those which are nasty, insulting, or unreasonable, and are probable to drive other users away from a discourse. The categorization of toxic comments is a subtask of sentiment analysis.

Severe Toxic: Adverse consequences that follow repeated or continuous ingestion of a test sample for a large amount of a person's life are stated to as Severe Toxic.

Obscene: When you say something is obscene, you're expressing that it offends you because it contains sex or violence in an inappropriate or unsettling way.

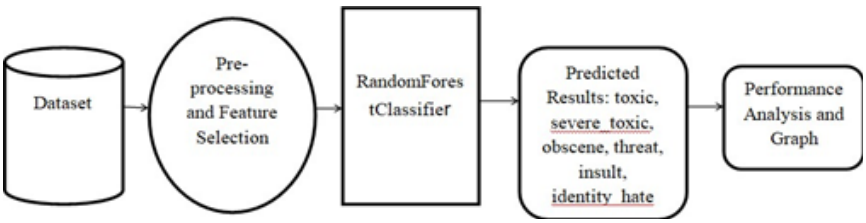

Figure 1 Proposed Architecture

Insult: An insult is a deliberately impolite deed or way of speaking, as well as a lack of regard, esteem or courteous behaviour.

Threat: It refers to a threatening message delivered in a terrible manner or a threatening message delivered in a horrifying way. Following pre-processing, we will instruct the model using the Random Forest Classifier. This notion is also referred to as predictive method. This approach may be used to prepare the model and is based on a dataset that has been coded or without instruction.

This tagged dataset is employed to train the random forest classifier model, It is used to categorize the comments into different levels.

our suggested system model is built with Random Forest Classifier, an algorithm for supervised learning that learns from training data and predicts the output for test data. This method produces a clustering.

Each feature has a more precise framework with less presumptions overfitting situations. It is the most widely used algorithm and is utilised in all ML-related real-time problems. It consists of a number of decision trees drawn from subsets of the datasets, with the average used to increase model accuracy. It also works with bigger datasets with high dimensionality. We created the system with the Flask web framework, and the user is prompted to submit a remark. The mentioned values are sent into the Random Forest Classifier

Based on the aforementioned input variables, our algorithm predicts the harmful class and displays the toxic level to the user. Our suggested system model's accuracy is typically extremely good.

Its efficiency is especially noticeable with large data sets. Provides an estimate of crucial categorization variables. The generated forests can be stored and reused. Unlike other models, it does not overburden itself with features. It gives an efficient method of dealing with missing data. Noise has less of an influence on Random Forest. Random Forest is typically resistant to outliers and can deal with them automatically.

Implementation

Data Collection

The Data Collection procedure is developed in the 1st module. Collecting data is the principal real step towards the actual construction of a (ML)machine learning model. It is a vital phase that will have a knock-on effect on how successful the model is; the greater and better data we have, the better our prototype will perform.

There are many methods for gathering data, including online scraping and manual interventions. Our dataset may be found in the model folder. The dataset is from the well- known dataset repository kaggle. The dataset's URL is provided below.

Data Preparation

Gather and arrange data for training. Clean up everything that needs it (remove duplicates, fix errors, deal with missing numbers, normalization, data type conversions, and so on).

Random data to eliminate the impacts of the sequence in which we acquired and/or otherwise prepared our data.

Visual data to aid in the detection of meaningful correlations between variables or class imbalances, or do other exploratory analysis. Sets are classified as training and assessment sets.

Model Selection

The Random Forests Algorithm

Let's go out the algorithm in layman's words. Assume you want to go on a trip and you want to go somewhere you would love. So, how do you go about finding a place you'll like? You may conduct an internet search, read reviews on travel blogs and websites, or ask your friends. Assume you chose to question your pals and asked them about their previous trip experiences to various locations. Every buddy will give you some recommendations. You must now construct a list of the recommended locations. Then you invite them to vote (or choose one best spot for the vacation) from your list of suggested destinations. The location with the most votes will be your final pick for the trip.

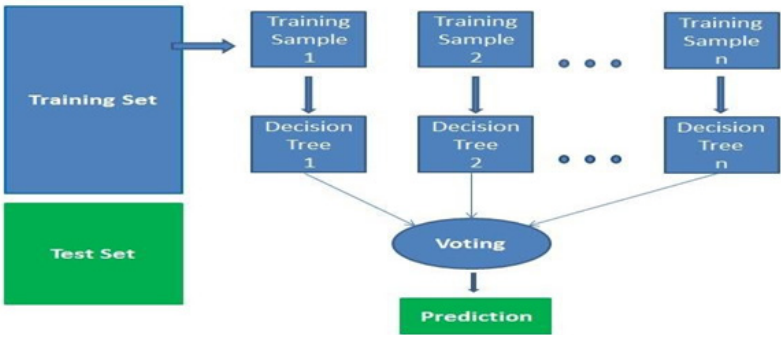

Figure 2 Model Selection

Results

Figure 4 Performance Analysis

Figure 4 shows performance analysis of the different data

Conclusion

In this study, we focused primarily on the remarks that are popular among teenage conversations on social media sites. The random forest Classifier was used to generate the model. We were able to categorise the comments into six distinct groups, and the random Forest Classifier produced more accurate findings. The model not only classifies a statement as toxic or non-toxic, but also gives percentages of Obscene, Toxic, Severe Toxic, Hate, Threat, and Identity Hate. Score(toxic)0.838055, Score (severe, toxic) 0.934874, Score(obscene) 0.909091, Score(insult) 0.883993, Score(threat) 0.795539, Score (identity, hate) 0.768448 were detected on the trained model. We demonstrated that our suggested approach is quite effective at detecting hazardous comments.

References

1. T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," arXiv, no.
2. P. L. Teh, C. Bin Cheng, and W. M. Chee, "Identifying and categorising profane words in hate speech," ACM International Conference Proceeding Series, pp. 65-69, 2018.
3. P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, 2018.
4. A. Gaydhani, V. Doma, S. Kendre, and L. Bhagwat, "Detecting hate speech and offensive language on Twitter with machine learning: An N-gram and TFIDF-based approach," 2018.

5. F. Del Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, F. Del Vigna, A. Cimino, F. Dell' Orletta, M. Petrocchi, and M."Hate me, hate me not: Hate speech detection on Facebook," CEUR Work shop Proc., vol. 1816, 2012.
6. V. Balakrishnan, S. Khan, T. Fernandez, and H. R. Arabnia, "*Cyberbullying detection on Twitter using big five and dark triad features," Personality and Individual Differences, vol. 141, no. 4, pp. 252-257, April 2019.
7. ''A deep learning strategy for automated hate speech identification in the Saudi Twittersphere,'' R. Alshalan and H. Al-Khalifa, 2020, Appl. Sci., vol. 10, no. 23, pp. 1-16.
8. B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, "'Measuring the Reliability of Hate Speech Annotations: The Case of B. Cabrera,"'
9. C. Ring, "Hate speech IN social media: An exploration of the problem and its proposed solutions," 2013.