Hate Speech Detection – the use of logistic regression.

Literature Review

Hate speech among people continues to grow, especially among the youths of the generation. To overcome the spread of hate Speech toward people of particular based on their identity, ethnicity and religion hate speech detection is important. The mechanism to identify and detect hate speech with the help of Natural Language Processing comes to an action and will further rise with the time. Prior research had also been done and numerous methods had been developed for the task including deep learning based methods. One of those methods includes the use of LSTM model (Long-short term memory) artificial neural network used in the field of artificial intelligence. My goal is to illuminate progress in the area and eliminate hate speech in the dataset. In order to create a hate speech detection algorithm, I am going to use python based NLP machine learning technique called logistic regression to train the algorithm to classify hate speech.

Abstract

Hate Speech is a very harmful activity which spreads hate towards a group or an individual member based on their identity or religion. Hate speech generally comes from social network domain because most people put up hate speech comments on certain social networks. Hate is a dangerous harmful activity which has been put up in recent times. It's detection is really important because it can help to reduce it, with the help of Natural Language Processing these problems can be identified and can be terminated accordingly. Automated hate speech detection is an important tool in combating the spread of hate speech in social media. The paper attempts to propose a model with the help of Natural Language Processing and machine learning technique called logistic regression, the algorithm is to classify hate speech using the data extracted. Using NLP technique, I will try to extract keywords that convey importance within hate speech.

Keywords

Hate speech detection, machine learning, Natural Language Processing, hate speech detecting algorithm, model to predict hate speech, classifying sexual harassment using machine learning, predicting the toxicity of comments, text classification, detecting hate tweets, twitter sentiment analysis, logistic regression.

Introduction

Hate Speech toward people of particular or religion is spreading in social media, this online hatred can lead to real life consequences such as fear and hate throughout communities. Hate speech detection is important because it can help prevent these harmful effects. Machine learning algorithms can be used to detect hate speech. These algorithms can analyze text and identify hate speech. They can also be used to determine the tone of a text. This can be used to identify hate speech that is disguised as jokes

or sarcasm. Hate speech generally comes from social network domain because most people put up hate speech comments on certain social networks, and that I will try to resolve in this project.

The scope of this project includes what are the different forms of hate they talk about and what is a basic social network and from where the dataset has been extracted. Machine leaning is a type of artificial intelligence that can be used to learn from data. It can be used to find patterns in data. Machine learning algorithms can be used to detect speech. These algorithms can analyze text and identify hate speech. They can also be used to determine the tone of a text.

Automated hate speech detection is an important tool in combating the spread of hate speech in social media. The techniques for detecting hate speech using machine learning include classifiers and deep learning. It can be used to find patterns in data. Natural Language processing techniques can be used to detect hate speech.

Students can learn how algorithms are working behind and create their own machine learning algorithms to further classify and detect hate speech. Faculty/Staff members can further enhance the existing algorithms or create their own models which will lead to classifying hate speech and detect hate speech. Cyber Security experts can use such algorithms or models to detect and identify those people spreading hatred and take over their accounts, ban from social sites and if most severe take them into custody. Game Developers can use these algorithms to target such account spreading toxic comments during the game and ban them or putting restrictions over their accounts.

There are some assumptions that need to be followed which are one should have knowledge or blacklist kind of stuff, collection of such dataset related to hate speech. Hate verb lists. Hateful terms and phrases for hate speech based on race, disability and so on from different web pages and sites. Abbreviations and variants of those words. There are some kinds of dictionaries which one can use so as to essentially rule based systems to detect hate speech very quickly in an efficient manner.

Some of the constraints in the project include datasets that need to be collected to train and test the model to predict and detect hate speech. Number of characters, words, and syllables in each sentence when it comes to detect hate speech. Abbreviations need to be identified properly. English level score is used such as if sentence looks much more formal then there is low probability of it being hate speech. Some websites present extremely antisemitic ideologies and conclusions in a very scientific manner, there we can say it will be objectively presented things on hate speech.

This algorithms and model can be used individually or can be deployed over the servers to detect hate speech in social media, websites, online games while sending texts or messages. The algorithms and models run over the servers all the time and should be available to stay away from the toxic and hateful speech. The impact of downtime of these models can lead to many unfortunate events and can lead to catastrophic failures of the administration, people spreading hatred will spread more hateful comments thus impacting the individuals or groups who try to avoid from such events.

Due to the wide spread of hate speech over the internet there is a need to study automatic detection of hate speech. By automating its detection, the spread of hateful content can be reduced.

Contribution

Collect data for the training of automatic classifiers to detect hate speech for that some necessary operations need to be performed.

The data needs to be converted into its vector representation. The vectors representation can be generated using different methods such as bag of words, tf-idf, word embeddings.

With the progress in deep learning based embeddings, tools such as word2vec, glove, fasttext, methods have to be applied to obtain more expressive representations.

Tf-idf vectorization will extract keywords that convey importance within hate speech.

So that finally machine learning technique called logistic regression, will train the algorithm or model to classify hate speech using the data extracted from the dataset after preprocessing.

Database of datasets are required to identify hateful, aggressive, and insulting texts. A representative sampling of available training and evaluation public datasets needs to be collected.

Dependencies

Datasets.

Collection of data from various sources such as twitter, Instagram, websites.

Popular social network datasets such as

1. Twitter: Waseem et al.2018

2. Twitter: Wijesiriwardene et al. 2020

3. Davidson et al. 2017

Feature engineering.

Important features are extracted from the dataset to analyze the most predictable outcome from the dataset.

Topic classification.

Using LDA (linear discriminant analysis) the most popular topic modelling technique right topics could be used as features.

Sentiment Analysis/classification.

Identifying opinions in text and labelling them as positive, negative, or neutral based on the emotions users express over the sites.

Classification

Frequency of personal pronouns, presence of emoticons.

Flesh reading escores.

Measure how correctly the grammar is used in the sentence.

Have some word embeddings or words being represented.

Using distributed representations of words to learn their vectorized representation which are used in text mining operations.

Deep learning

Traditional approaches like TF-IDF, recently developed word embedding. Convolutional neural networks (CNN), Long-Short term memory (LSTM) are the three most popular deep neural network designs used for hate speech detection using deep learning models.

Packages such as NumPy, pandas, matplotlib, pyplot, Wordcloud, sklearn.

NLP tools: re, nltk.

Leveraging metadata.

Data Augmentation.

Handling Bias.

Multilabel classification.

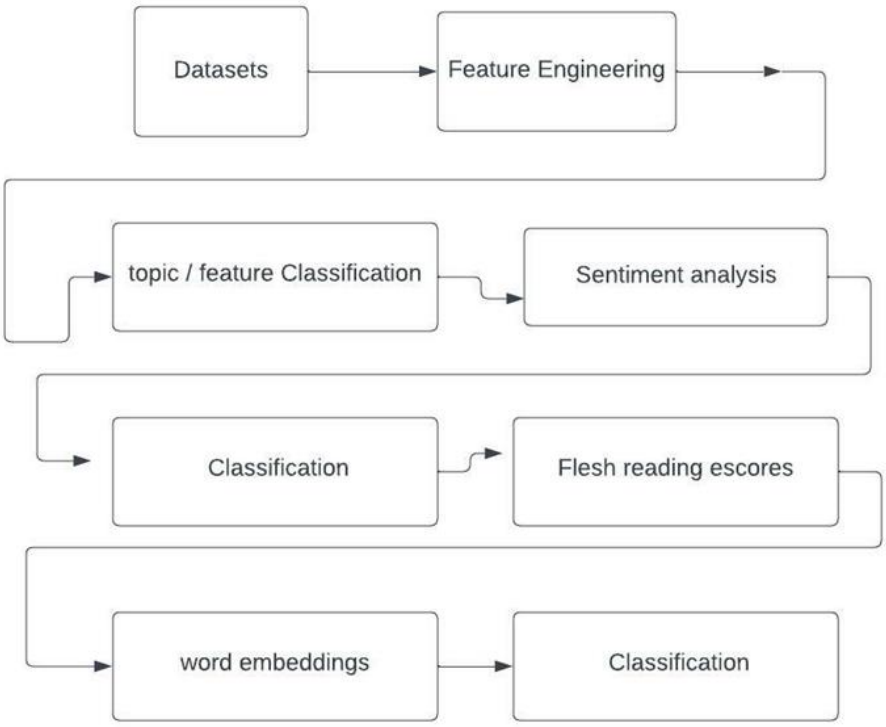
Classifiers/Regressors –

SVMs

Logistic regression

Random forests MLPs

Naïve Bayes.



Methodology

In order to create a hate speech detecting algorithm, I am going to use python based NLP machine learning techniques. Machine learning is basically teaching machines to accomplish various tasks by training them through data.

First collection of data is important, can get the data from anywhere and import it over the platform.

Then using a NLP (Natural language Processing) technique called Tf-Idf vectorization, I will extract keywords that convey importance within hate speech. This process is done by assigning Tf-idf scores to each word in a document.

Converting texts into numbers so the algorithm can understand it. By focusing on the words that convey importance, extract keywords from documents in order to understand it.

Finally, based on machine learning technique called logistic regression, I will train the algorithm to classify hate speech using the data extracted, however machine learning algorithms that can be used to detect hate speech include Naïve Bayes, Support Vector machines, extreme gradient boosting, multi-layer perception, long – short term memory networks.

In order to train the classifier, the data needs to be converted into its vector representation. The vectors representation can be generated using different methods such as bag of words, Tf-

idf, and word embeddings. With the progress in deep learning based embeddings, tools such as word2vec, glove, fasttext, and transform based methods can be applied to obtain more expressive representations.

Import libraries required for solving and in detection of hate speech.

Import tools for Natural Language Processing.

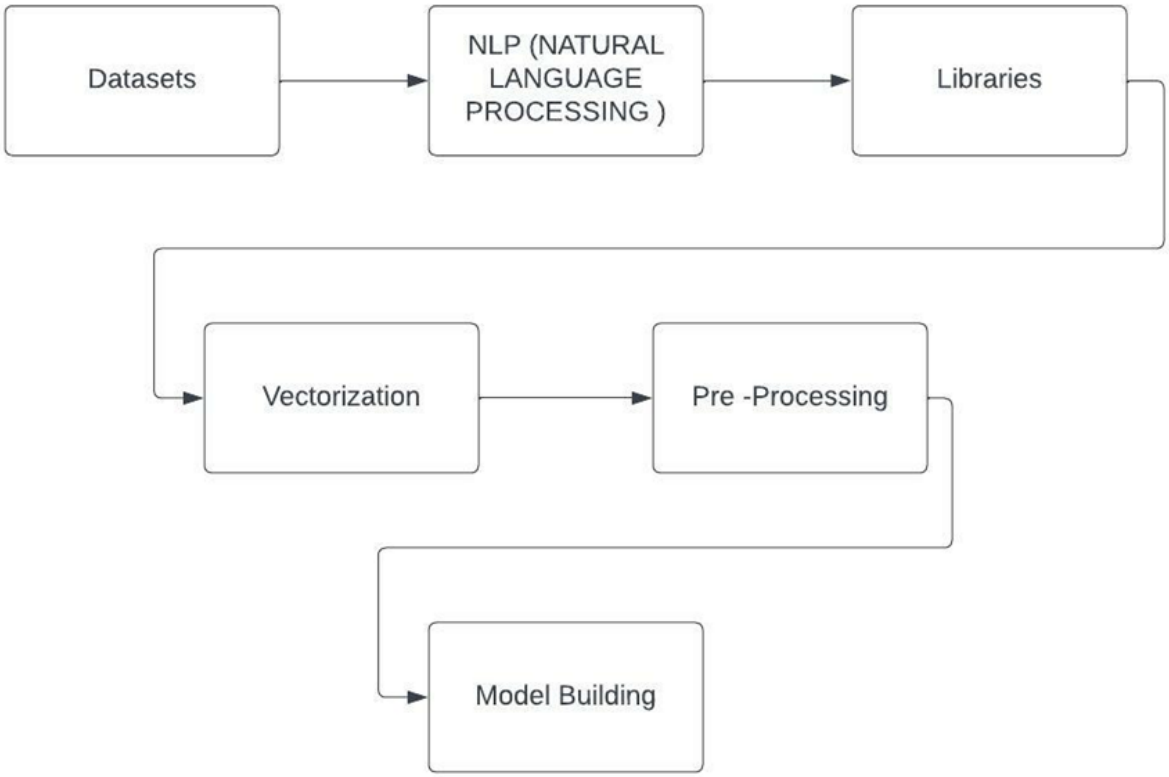
Import models for the dataset.

Import the dataset that is required for analysis.

Split the data into train/test dataset.

Some preprocessing is required in the dataset.

Finally applying the model and drawing a confusion matrix out of it.



Result Analysis

The hate speech detection datasets are usually not clean, so they need to be preprocessed before classification algorithms can be applied and that can detect hate speech in them.

Different machine learning models have different strengths that make some algorithms better than others for certain tasks such as detecting hate speech. Some models are more accurate while others are more efficient.

It is important to use different models and compare their performance in order to find the best one for hate speech detection. Pre training models have become popular in recent times and it is important to test whether they work well with hate speech detection algorithms.

It is also important to see how hate speech detection models can be used to address domain changes.

I tried to identify and examine challenges faced by online approaches for hate speech detection in text. Among these difficulties are subtitles in language, differing definitions, limitations of data availability for training and testing of these systems.

The algorithms or the models need to be modified or changed in accordance with the time in order to provide high reliability on the system, provide real time results, update the word cloud, bag of words, word embeddings, sentiment analysis should be done to classify the results so they need to be preprocessed before classification algorithms can be applied and that can detect hate speech in them.

As I have used a machine learning technique called logistic regression, I have trained the algorithm to classify hate speech using the data extracted. In accordance to that my model gives as accuracy of 0.90 which is a pretty decent score.

Conclusion

Hate speech is an important problem, websites that allow users to leave feedback, having negative impact on their online business and overall user experience. Now these algorithms or models can be deployed and can be triggered whenever there is hated speech thus leading to removal of those comments or the feedback or strict action against those accounts.

Hate speech and toxic comments in the chats are very common among the youths which they pass on among each other during a round or match of a game. These need to be controlled or tackled effectively in order to maintain a healthy environment among the players of each side.

Deep learning is a type of machine learning that can be used to learn from data. It can be used to find patterns in data. Machine learning algorithms can be used to detect hate speech.

These algorithms can analyze text and identify hare speech. They can also be used to determine the tone of a text. This can be used to identify hate speech that is than being classified has hate speech or not.

Automated hate speech detection is an important tool in combating the spread of hate speech, on social media sites, online gaming, or websites.

References

https://vitalflux.com/hate-speech-detection-using-machine-learning/

https://towardsdatascience.com/my-first-machine-learning-project-designing-a-hate-speech-detecting-algorithm-56ab32f10833

https://www.analyticsvidhya.com/blog/2022/02/classifying-sexual-harassment-using-machine-learning/

https://www.analyticsvidhya.com/blog/2022/07/predicting-the-toxicity-of-comments-using-text-classification/

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221152

https://arxiv.org/pdf/1803.03662.pdf

https://towardsdatascience.com/detecting-hate-tweets-twitter-sentiment-analysis-780d8a82d4f6

https://ai.facebook.com/blog/how-facebook-uses-super-efficient-ai-models-to-detect-hate-speech/

https://about.instagram.com/blog/announcements/an-update-on-our-work-to-tackle-abuse-on-instagram

https://dl.acm.org/doi/abs/10.1145/2740908.2742760