Hate-speech detection on social media

Machine Learning and Artificial Intelligence
Berkely Eng, Berkely Hass. Nov 2022.

Hagay Zamir
June 7, 2023

Abstract

This project focuses on the challenges faced in automatic hate-speech detection on social media, particularly the difficulty in distinguishing hate speech from other forms of offensive language. Traditional lexical detection methods tend to have low precision, as they classify any messages containing specific terms as hate speech. Previous supervised learning approaches also struggled to differentiate between the two categories effectively. To address this, a crowd-sourced hate speech lexicon was utilized to collect tweets containing hate speech keywords. A multi-class classifier was trained using this dataset, and close analysis of the predictions revealed instances where hate speech could be reliably separated from other offensive language, as well as situations where differentiation proved more challenging. The findings indicated that racist and homophobic tweets were more likely to be classified as hate speech, while sexist tweets were predominantly labeled as offensive. Tweets lacking explicit hate keywords posed additional difficulties in classification. This project provides valuable insights into the complexities of hate-speech detection and underscores the significance of thorough analysis for enhancing classification accuracy.

Introduction

Defining hate speech and differentiating it from offensive language is a complex task without a formal consensus. Generally, hate speech refers to speech that targets disadvantaged social groups in a potentially harmful manner. In the United States, hate speech is protected under the First Amendment but has been extensively debated in legal and campus speech code contexts. Other countries like the United Kingdom, Canada, and France have laws prohibiting hate speech, often defined as speech that targets minority groups and could incite violence or

social disorder, carrying penalties such as fines and imprisonment. Online platforms like Facebook and Twitter have implemented policies to curb hate speech by prohibiting attacks based on characteristics like race, ethnicity, gender, and sexual orientation, or threats of violence. Our definition of hate speech encompasses language expressing hatred, derogation, humiliation, or insults towards targeted groups, including extreme cases that involve threats or incitement to violence. However, our definition does not encompass all instances of offensive language, as certain terms may be used differently within specific communities without the same intent of hate. To address the conflation of hate speech and offensive language, we labeled tweets into three categories: hate speech, offensive language, or neither. Using this dataset, we trained a model to distinguish between these categories and conducted a detailed analysis to understand the challenges involved in accurate classification. The results highlight the importance of fine-grained labeling and the need to consider context and the diverse usage of hate speech in future research.

Related Work

Bag-of-words approaches, while having high recall, often result in high false positive rates when classifying tweets as hate speech due to the presence of offensive words. This challenge is particularly pronounced given the frequent occurrence of offensive language and curse words on social media. Research focused on anti-black racism found that a significant portion of tweets categorized as racist were labeled as such primarily because they contained offensive words. Distinguishing between hate speech and other forms of offensive language relies on subtle linguistic nuances. For instance, tweets containing the word "ngger" are more likely to be classified as hate speech compared to "ngga." Ambiguities arise,

such as the word "gay" being used pejoratively or in unrelated contexts. Syntactic features have been explored to improve hate speech identification, such as identifying relevant noun-verb pairs or employing specific POS trigrams. However, many supervised approaches have conflated hate speech with offensive language, making it challenging to accurately identify hate speech instances. Although neural language models show promise, existing training data often lack a precise definition of hate speech. Integrating non-linguistic features like author gender or ethnicity could enhance hate speech classification, but such information is frequently unavailable or unreliable on social media platforms.

Dataset Overview and Data Preparation

Introduction:

This section provides an overview of the dataset used for the final paper and outlines the data preparation steps undertaken before analysis. The dataset consists of tweets collected from Twitter, which were manually coded for hate speech and offensive language. Additionally, statistical information about the dataset's columns and data types is presented.

Dataset Overview:

The dataset used for the final paper comprises a total of 24,783 rows and 7 columns. These columns include 'count', 'hate_speech', 'offensive_language', 'neither', 'class', 'tweet', and 'processed_tweet'. The 'count' column represents the total number of people who rated each tweet, while 'hate_speech', 'offensive_language', and 'neither' columns indicate the number of raters who classified a tweet as hate speech, offensive language, or neither, respectively. The

'class' column assigns a label to each tweet (0 for offensive, 1 for hate speech, and 2 for neither).

Data Types and Missing Values:

The dataset consists of integer and object data types. The integer columns ('count', 'hate_speech', 'offensive_language', 'neither', and 'class') hold numerical information, while the object columns ('tweet' and 'processed_tweet') contain text data. Importantly, no missing values are present in any of the columns, as denoted by a count of 0 for each column.

Descriptive Statistics:

Basic statistics are provided for the continuous columns. The 'count' column has a mean of 3.243 and a standard deviation of 0.883, with a minimum of 3 and a maximum of 9. The 'hate_speech' column has a mean of 0.281, a standard deviation of 0.632, and ranges from 0 to 7. The 'offensive_language' column has a mean of 2.414, a standard deviation of 1.399, and ranges from 0 to 9. The 'neither' column has a mean of 0.549, a standard deviation of 1.113, and ranges from 0 to 9. The 'class' column, representing the assigned labels, has a mean of 1.110, a standard deviation of 0.462, and ranges from 0 to 2.

Categorical Distribution:

Within the 'class' column, there are three categories: 0, 1, and 2. Class 1 has the highest occurrence, with 19,190 instances, followed by Class 2 with 4,163 instances and Class 0 with 1,430 instances.

Additional Statistics:

Additional statistics are provided for each continuous column. The 'count' column has a mean of 3.243, a median of 3, and a mode of 3. The 'hate_speech' column has a mean of 0.281, a median of 0, and a mode of 0. The 'offensive_language' column has a mean of 2.414, a median of 3, and a mode of 3.

The 'neither' column has a mean of 0.549, a median of 0, and a mode of 0. These statistics offer further insights into the distribution and central tendencies of the data.

Conclusion:

The dataset used in the final paper provides a comprehensive collection of tweets coded for hate speech and offensive language. Statistical summaries of the dataset's columns help understand the distribution and characteristics of the data. With the data prepared and the dataset overview established, the subsequent sections of the paper can delve into the analysis and modeling processes.

Models

This project aims to explore the performance of various machine learning models for hate speech and offensive language classification. The models considered include Logistic Regression, Decision Tree, Random Forest, SVC, K Neighbors, and XGBoost, combined with different vectorizers such as Count, Tfidf, Hashing, and Binary. The evaluation of these models is based on two key metrics: 'best_score' representing accuracy and 'fit_time' indicating training time.

Among the models analyzed, the SVC Tfidf Vectorizer achieves the highest 'best_score' of 90.64% but requires a considerable training time of approximately 126.92 minutes. On the other hand, the Decision Tree Hashing Vectorizer achieves a slightly lower 'best_score' of 89.08% but offers a significantly shorter training time of around 1.53 minutes, making it more suitable for efficient computations.

Additionally, the baseline models, Naive Bayes and Logistic Regression, exhibit reasonable accuracy with 'best_score' values of 86.91% and 89.75%, respectively, and remarkably fast training times of 0.0024 and 0.0127 minutes. These models are particularly advantageous for applications that require rapid training.

Moreover, the eXtreme Gradient Boosting model achieves a 'best_score' of 91.57% with a training time of 0.915651 minutes, while the AdaBoost model achieves a 'best_score' of 90.11% with a training time of 0.380739 minutes. The Random Forest model attains a 'best_score' of 89.15% with a training time of 5.282936 minutes.

Results

The results of our study demonstrate the performance of various machine learning models combined with different vectorizers for hate speech and offensive language classification. We evaluated models such as Logistic Regression, Decision Tree, Random Forest, SVC, K Neighbors, and XGBoost, using metrics like 'best_score' for accuracy and 'fit_time' for training time.

Among the models tested, the SVC Tfidf Vectorizer achieved the highest 'best_score' of 90.64% but required a substantial training time of approximately 126.92 minutes. In contrast, the Decision Tree Hashing Vectorizer achieved a slightly lower 'best_score' of 89.08% but exhibited a significantly shorter training time of approximately 1.53 minutes, making it more efficient for real-time computations.

The baseline models, Naive Bayes and Logistic Regression, demonstrated reasonable accuracy with 'best_score' values of 86.91% and 89.75% respectively. These models also exhibited remarkably fast training times of 0.0024 and 0.0127 minutes, making them suitable for applications requiring rapid training.

Additional results showed that the eXtreme Gradient Boosting model achieved a 'best_score' of 91.57% with a training time of 0.915651 minutes, while the AdaBoost model achieved a 'best_score' of 90.11% with a training time of 0.380739

minutes. The Random Forest model achieved a 'best_score' of 89.15% with a training time of 5.282936 minutes.

Conclusion

In conclusion, our study focused on the challenges of automatic hate-speech detection and offensive language classification on social media. We explored various machine learning models combined with different vectorizers to tackle these challenges. The results demonstrated the performance of models such as Logistic Regression, Decision Tree, Random Forest, SVC, K Neighbors, and XGBoost.

The SVC Tfidf Vectorizer achieved the highest accuracy ('best_score') of 90.64% but required a significant training time. On the other hand, the Decision Tree Hashing Vectorizer achieved a slightly lower accuracy of 89.08% but had a much shorter training time, making it more suitable for real-time applications. The baseline models, Naive Bayes and Logistic Regression, displayed reasonable accuracy and fast training times, making them practical choices for applications with time constraints.

Additional results highlighted the effectiveness of the eXtreme Gradient Boosting and AdaBoost models, which achieved high accuracies of 91.57% and 90.11%, respectively, with relatively short training times. The Random Forest model also showed good accuracy at 89.15%, albeit with a longer training time.

These findings emphasize the importance of selecting the appropriate model based on the desired trade-off between accuracy and computational efficiency. The study contributes valuable insights into hate-speech detection and offensive language classification, aiding in the development of robust and efficient models

for these tasks. Future research can build upon these results to further enhance classification accuracy and address the complexities associated with distinguishing hate speech from other forms of offensive language in online platforms.