Hate-speech detection on social media

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

Gour Gopal Nandi
October 8, 2023

Abstract

This project primarily focuses on the difficulties encountered in automatic hate-speech detection on social media, particularly the challenge of distinguishing hate speech from other forms of offensive language. Conventional lexical detection methods often suffer from low precision, categorizing any messages containing specific terms as hate speech. Prior supervised learning approaches have also faced challenges in effectively distinguishing between these two categories. To tackle this issue, we employed a crowd-sourced hate speech lexicon to gather tweets that included hate speech keywords. Using this dataset, we trained a multi-class classifier, and a detailed examination of the predictions highlighted situations where hate speech could be reliably separated from other offensive language, as well as cases where differentiation proved more intricate. The findings revealed that racist and homophobic tweets were more likely to be classified as hate speech, whereas sexist tweets were predominantly labeled as offensive. Tweets without explicit hate keywords presented additional complexities in terms of classification. This project offers valuable insights into the nuances of hate-speech detection and underscores the importance of thorough analysis in improving classification accuracy.

Introduction

Defining hate speech and discerning it from offensive language is a complex undertaking without a formal consensus. In general, hate speech pertains to speech that targets disadvantaged social groups in a potentially harmful manner. In the United States, hate speech is safeguarded by the First Amendment but has been extensively debated in legal and campus speech code contexts. Contrastingly, other countries like the United Kingdom, Canada, and France have enacted laws prohibiting hate speech, often characterized as speech targeting

minority groups that could incite violence or social disorder, carrying penalties such as fines and imprisonment.

Online platforms such as Facebook and Twitter have implemented policies to mitigate hate speech by prohibiting attacks based on characteristics like race, ethnicity, gender, and sexual orientation, as well as threats of violence. Our definition of hate speech encompasses language expressing hatred, derogation, humiliation, or insults towards targeted groups, including extreme cases involving threats or incitements 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 categorized tweets into three groups: hate speech, offensive language, or neither. Leveraging this dataset, we trained a model to differentiate between these categories and conducted an in-depth analysis to comprehend the challenges associated with precise classification. The results underscore the importance of nuanced labeling and emphasize the necessity of considering context and the diverse usage of hate speech in future research.

Related Work

Bag-of-words methods, while achieving high recall, frequently yield elevated false positive rates when classifying tweets as hate speech due to the inclusion of offensive language. This challenge is especially noticeable given the frequent occurrence of offensive vocabulary and profanities on social media platforms. Research concentrated on addressing anti-black racism uncovered that a substantial portion of tweets categorized as racist were primarily labeled as such

due to the presence of offensive terms. Distinguishing between hate speech and other forms of offensive language hinges on subtle linguistic nuances. For example, tweets containing the word "ngger" are more prone to be categorized as hate speech compared to "ngga." Ambiguities can emerge, such as the term "gay" being used pejoratively or in unrelated contexts.

Syntactic features have been investigated to enhance the identification of hate speech, including the recognition of relevant noun-verb pairs or the utilization of specific POS trigrams. However, many supervised approaches have conflated hate speech with offensive language, posing challenges in accurately identifying instances of hate speech. While neural language models hold promise, existing training data often lack a precise definition of hate speech. The incorporation of non-linguistic attributes like author gender or ethnicity could potentially improve hate speech classification, but such information is frequently unavailable or unreliable on social media platforms.

Dataset Overview and Data Preparation

Introduction:

This section offers an introduction to the dataset utilized in the final paper and delineates the steps taken in preparing the data for analysis. The dataset comprises tweets gathered from Twitter, which underwent manual coding to identify

instances of hate speech and offensive language. Furthermore, the section presents statistical details regarding the dataset's columns and data types.

Dataset Overview:

The dataset utilized in the final paper encompasses a grand total of 24,783 rows distributed across 7 columns. These columns are denoted as 'count,' 'hate_speech,' 'offensive_language,' 'neither,' 'class,' 'tweet,' and 'processed_tweet.' Within this dataset, the 'count' column signifies the cumulative count of individuals who assessed each tweet. Meanwhile, the 'hate_speech,' 'offensive_language,' and 'neither' columns indicate the number of evaluators who categorized a tweet as hate speech, offensive language, or neither, respectively. The 'class' column assigns a categorical label to each tweet, with '0' representing offensive, '1' representing hate speech, and '2' indicating neither.

Data Types and Missing Values:

The dataset encompasses both integer and object data types. The integer columns ('count,' 'hate_speech,' 'offensive_language,' 'neither,' and 'class') contain numerical data, while the object columns ('tweet' and 'processed_tweet') house textual information. Notably, all columns are devoid of missing values, as indicated by a count of 0 for each column.

Descriptive Statistics:

Summary statistics are provided for the continuous columns. The 'count' column exhibits an average of 3.243, with a standard deviation of 0.883. Its values range from a minimum of 3 to a maximum of 9. The 'hate_speech' column displays an average of 0.281, a standard deviation of 0.632, and values ranging from 0 to 7. The 'offensive_language' column shows an average of 2.414, a standard deviation of 1.399, and values spanning from 0 to 9. The 'neither' column reports an average of 0.549, a standard deviation of 1.113, and values ranging from 0 to 9. Regarding the

'class' column, which signifies assigned labels, it has an average of 1.110, a standard deviation of 0.462, and values spanning from 0 to 2.

Categorical Distribution:

Within the 'class' column, three categories exist: 0, 1, and 2. Class 1 exhibits the highest frequency, with 19,190 instances, followed by Class 2 with 4,163 instances, and Class 0 with 1,430 instances.

Additional Statistics:

Additional statistics are furnished for each continuous column. The 'count' column displays an average of 3.243, a median of 3, and a mode of 3. The 'hate_speech' column demonstrates an average of 0.281, a median of 0, and a mode of 0. The 'offensive_language' column showcases an average of 2.414, a median of 3, and a mode of 3. Finally, the 'neither' column presents an average of 0.549, a median of 0, and a mode of 0. These statistics provide further insights into the distribution and central tendencies of the data.

Conclusion:

The dataset employed in the final paper offers a comprehensive compilation of tweets that have been categorized for hate speech and offensive language. Statistical summaries of the dataset's columns aid in comprehending the data's distribution and attributes. With the data now prepared and the dataset overview in place, the subsequent sections of the paper can delve into the analysis and modeling processes.

Models

The objective of this project is to explore the performance of various machine learning models for the classification of hate speech and offensive language. The models under consideration include Logistic Regression, Decision Tree, Random

Forest, Support Vector Classifier (SVC), K Neighbors, and XGBoost, in combination with different vectorization techniques like Count, Tfidf, Hashing, and Binary. The assessment of these models is based on two primary metrics: 'best_score,' which represents accuracy, and 'fit_time,' indicating the time taken for training.

Among the models examined, the SVC Tfidf Vectorizer achieves the highest 'best_score' of 90.64%. However, it demands a substantial training time of approximately 126.92 minutes. Conversely, the Decision Tree Hashing Vectorizer achieves a slightly lower 'best_score' of 89.08%, but it offers a significantly shorter training time of around 1.53 minutes, rendering it more suitable for efficient computations.

Furthermore, the baseline models, Naive Bayes and Logistic Regression, demonstrate respectable accuracy, with 'best_score' values of 86.91% and 89.75%, respectively. These models exhibit remarkably fast training times, with Naive Bayes at 0.0024 minutes and Logistic Regression at 0.0127 minutes. These models prove advantageous for applications that necessitate rapid training.

Additionally, the eXtreme Gradient Boosting model achieves a 'best_score' of 91.57% with a training time of 0.915651 minutes, while the AdaBoost model attains a 'best_score' of 90.11% with a training time of 0.380739 minutes. The Random Forest model achieves a 'best_score' of 89.15% with a training time of 5.3 minutes.

Results

My study's findings provide insights into the performance of various machine learning models in conjunction with different vectorization techniques for the classification of hate speech and offensive language. We assessed models such as Logistic Regression, Decision Tree, Random Forest, Support Vector Classifier (SVC),

K Neighbors, and XGBoost, using evaluation metrics like 'best_score' for accuracy and 'fit_time' for training duration.

Among the models examined, the SVC Tfidf Vectorizer achieved the highest 'best_score' at 90.64%, albeit at the cost of a significant training time of approximately 126.92 minutes. Conversely, the Decision Tree Hashing Vectorizer achieved a slightly lower 'best_score' of 89.08% but demonstrated a notably shorter training time of about 1.53 minutes, rendering it more efficient for real-time computations.

The baseline models, Naive Bayes and Logistic Regression, exhibited commendable accuracy, with 'best_score' values of 86.91% and 89.75%, respectively. These models also showcased exceptionally swift training times, with Naive Bayes taking 0.0024 minutes and Logistic Regression taking 0.0127 minutes. These characteristics make them well-suited for applications requiring rapid model training.

Furthermore, our additional results revealed that the eXtreme Gradient Boosting model attained 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 5.3 minutes.

Conclusion

In conclusion, our study centered on the challenges of automatic hate-speech detection and offensive language classification within the realm of social media. We explored a range of machine learning models in conjunction with different

vectorization techniques to address these challenges. The results shed light on the performance of various models, including Logistic Regression, Decision Tree, Random Forest, SVC, K Neighbors, and XGBoost.

The SVC Tfidf Vectorizer emerged as the top performer, achieving an accuracy ('best_score') of 90.64%. However, it necessitated a substantial training time. In contrast, the Decision Tree Hashing Vectorizer delivered a slightly lower accuracy at 89.08% but boasted a significantly shorter training duration, rendering it well-suited for real-time applications. The baseline models, Naive Bayes and Logistic Regression, demonstrated respectable accuracy and rapid training times, making them practical choices for time-sensitive applications.

Furthermore, our supplementary findings underscored the effectiveness of the eXtreme Gradient Boosting and AdaBoost models, which achieved high accuracies of 91.57% and 90.11%, respectively, with relatively brief training periods. The Random Forest model also exhibited commendable accuracy at 89.15%, albeit with a lengthier training time.

These insights underscore the importance of selecting an appropriate model that aligns with the desired balance between accuracy and computational efficiency. The study contributes valuable insights to the field of hate-speech detection and offensive language classification, aiding in the development of robust and efficient models for these critical tasks. Future research can leverage these findings to further enhance classification accuracy and address the intricate nuances associated with differentiating hate speech from other forms of offensive language on online platforms.