Detection of Hate Speech in Text Sentences

Nishit Rathod, Ahmed Alshraideh, Dhyey Chauhan

Analysis Report:
Detection of Hate Speech in Text Sentences

Introduction:

The objective of this analysis is to detect hate speech in text sentences using variants and approximations of Support Vector Machines (SVM). Hate speech detection is an important task in online platforms to ensure a safe and inclusive environment for users. The analysis will involve training SVM models on a provided training dataset and evaluating their performance on a test dataset.

Dataset Description:

- The training dataset (train.csv) contains comments along with their binary labels indicating whether they are toxic or not.
- The test dataset (test.csv) is used to predict the toxicity probabilities for the comments.
- The test_labels.csv file provides labels for the test data, where a value of -1 indicates that the comment was not used for scoring.
- The sample_submission.csv file is a sample submission file in the correct format.

Data Preprocessing:

- The provided code snippet includes several preprocessing steps such as removing punctuation, converting text to lowercase, tokenizing words, removing stopwords, and applying TF-IDF transformation to convert text into numeric vectors.
- These preprocessing steps are essential to clean the text data and convert it into a suitable format for training machine learning models.

Feature Extraction:

- The code snippet also includes the generation of word embeddings using Word2Vec and Doc2Vec models.
- Word2Vec model learns word embeddings based on the context of words, while Doc2Vec model learns document embeddings that capture the semantic meaning of the sentences.
- These embeddings represent the textual information in a dense vector space, which can be used as features for training SVM models.

Training SVM Models:

- The code snippet trains SVM models using different sets of features, including average word embeddings, TF-IDF weighted average word embeddings, and Doc2Vec embeddings.

- SVM models are trained using the LinearSVC and SVC (kernel='rbf') implementations.
- The trained SVM models are saved using pickle for future use.

Evaluation and Findings:

- The trained SVM models are evaluated on the test dataset to assess their performance in detecting hate speech.
- Classification reports are generated, which include metrics such as precision, recall, F1-score, and support for each class (toxic and non-toxic).
- The performance of the models is assessed based on these metrics, providing insights into their effectiveness in identifying hate speech.

Recommendations:

- Based on the evaluation results, the SVM model using Doc2Vec embeddings achieved the highest performance in detecting hate speech.
- This model should be considered for deployment in the online platform to identify and take appropriate actions against hate speech.
- Further analysis and fine-tuning of the models can be done to improve their performance, such as hyperparameter tuning, ensemble methods, or exploring different text representation techniques.
- Regular updates and retraining of the models should be carried out to adapt to evolving patterns of hate speech and maintain effective detection.

Conclusion:

- Hate speech detection is a challenging task, but by employing variants and approximations of Support Vector Machines along with appropriate feature extraction techniques, it is possible to achieve accurate results.
- The analysis highlights the importance of creating safe and inclusive online environments by proactively identifying and addressing hate speech.
- The provided code snippet serves as a starting point for hate speech detection, and further enhancements can be made based on the specific requirements and characteristics of the target platform.