INFORMATION RETRIEVAL AND SEMANTIC WEB

HATE SPEECH DETECTION

PROJECT REPORT

SUBMITTED TO:
DR NEETU SARDANA

SUBMITTED BY:
DEVANSHI KAPLA 20103176 B6
ADITYA JAMWAL 20103283 B10
TANUPRIYA PATHAK 20103288 B10

Objective

The objective of this task is to detect hate speech in tweets. For the sake of simplicity, we say a tweet contains hate speech if it has a racist or sexist sentiment associated with it. So, the task is to classify racist or sexist tweets from other tweets.

Approach

We have followed a sequence of steps needed to solve a general sentiment analysis problem. We will start with preprocessing and cleaning of the raw text of the tweets. Then we will explore the cleaned text and try to get some intuition about the context of the tweets. After that, we will extract numerical features from the data and finally use these feature sets to train models and identify the sentiments of the tweets.

Feature Extraction

Two feature extraction techniques have been utilised in this project:

Bag of Words (BoW)

The Bag of Words approach converts text into numerical features. Each tweet is represented as a vector, where each element corresponds to the count of a specific word in the tweet. The scikit-learn CountVectorizer is used to implement BoW.

TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF is another technique used to convert text into numerical features. It assigns weights to words based on their frequency in a tweet and inverse document frequency across all tweets. The scikit-learn TfidfVectorizer is used to implement TF-IDF.

Models Implemented

Four machine learning models have been implemented for sentiment analysis:

1. Logistic Regression

Logistic Regression is a simple yet effective binary classification algorithm. In this project, it is adapted for multi-class classification by using the "one-vs-rest" approach.

2. Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting.

3. Support Vector Machine (SVM)
SVM is a powerful classification algorithm that finds the best hyperplane to separate classes in a high-dimensional feature space.

4. Naive Bayes:
Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem, assuming independence among features, commonly used for text classification and spam filtering.

5. K-Nearest Neighbors (KNN):
KNN is a non-parametric and lazy learning algorithm that classifies a data point based on the majority class among its k nearest neighbors in the feature space, making it effective for pattern recognition and regression tasks.

Conclusion

This Hate Speech Detection project demonstrates the effectiveness of Bag of Words and TF-IDF techniques for feature extraction, along with four different machine learning models for sentiment classification. The model tuning process enhances the models' performance and ensures better generalisation to new data.

References

[1] Akuma, S., Lubem, T., & Adom, I. T. (2022). Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets. *International Journal of Information Technology*, *14*(7), 3629-3635.

[2] Mugambi, S. K. (2017). *Sentiment analysis for hate speech detection on social media: TF-IDF weighted N-Grams based approach* (Doctoral dissertation, Strathmore University).