Hate-Speech Classification

AUTHORS: Dalia Jeiroudi, Divya Kuma | Fabio Costa, Mohsin Braer, Vinicius Valverde

INTRODUCTION

- Exponential growth of social media has brought with it an increasing propagation of hate speech.
- Ease of communication has led to an increase of exchange of ideas between people
  - How to navigate between freedom of speech and the threatening use of insulting and offensive language?
- Objective: Construct the best model that can detect hate-speech in order to wipe it off the internet before it does any damage.

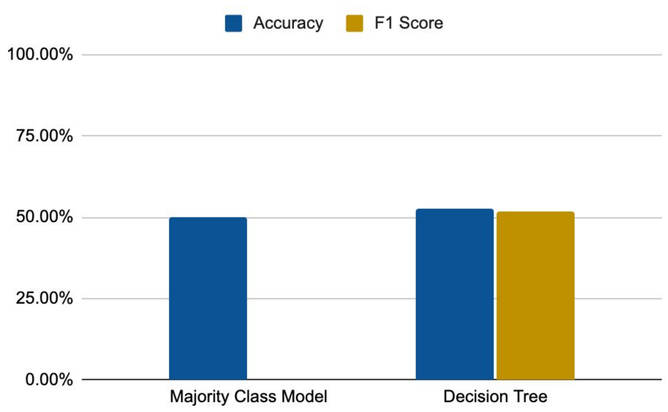
DATA

- Utilizing a dataset of ~25K tweets obtained from kaggle
  - Unbalanced dataset consisting of 5% of tweets categorized as hate
  - Removed 90% of data to make the dataset equal
- Cleaned data using NumPy, re, and NLTK to remove unnecessary twitter characters and downcase each tweet.
- Utilized datasets and transformers libraries to tokenize and preprocess data for our BERT model.

METHODOLOGY

- 80-20 train-test split using sklearn
- Trained the BERT model using transformers libraries.
- Used Sklearn libraries to train Logistic Regression, Naive-Bayes and K-NN classifiers. Used — neighbours
- Used Word2Vec embedding matrix to train multilayer perceptron
- Utilized Keras library to tokenize tweets & build Sequential model

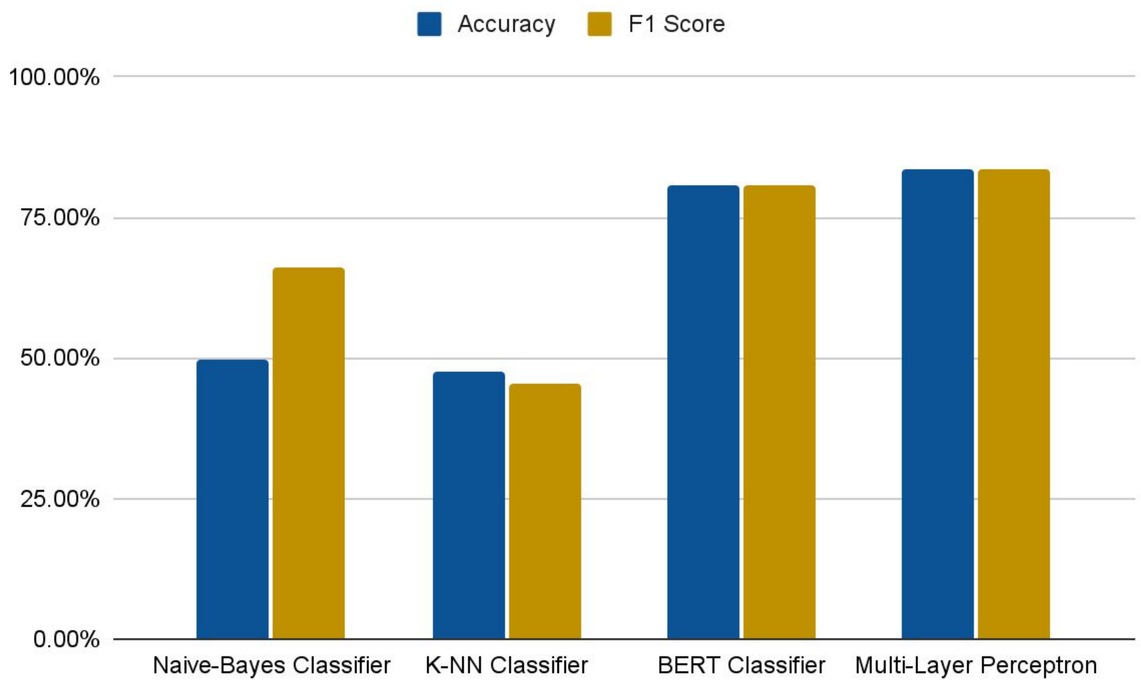
BASELINES



RESULTS

Model	Accuracy	F1 Score
Majority Class	0.5	0.0
Decision Tree	0.527	0.520
Naive-Bayes	0.498	0.660
K-NN	0.476	0.453
BERT	0.806	0.807
Multi-Layer Perceptron	0.836	0.836

ANALYSIS



CONCLUSION

- MLP and BERT models had highest accuracy rates at 0.823 and 0.806 respectively
- Multi-layer perceptron model had a high accuracy 0.836 and F1 score 0.836, therefore it is the best overall model
- In the future we would like to collect our own data in order to produce a more balanced dataset
- We could also consider other factors such as the author of the tweet