Hate Speech and Offensive Language Recognition

Natural Language Processing

Dimitris Patiniotis Spyropoulos,

July 1, 2022

Table of Contents

Introduction

Hate Speech and Offensive Language Recognition

- Practical solution to filter spam in content sharing apps/websites
- A multi-class classification problem
- Different ways to approach the problem

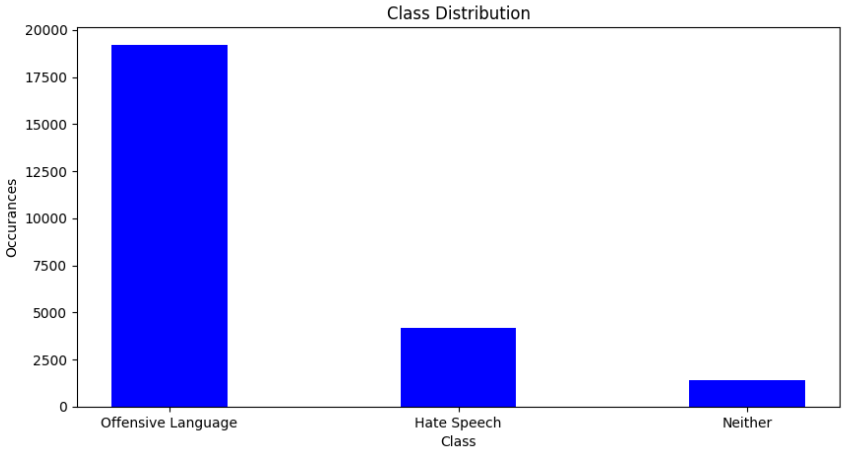
Methodology

Data Collection and Preparation

Dataset used from Davidson et al. (2017).

`https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset?datasetId=723100&sortBy=voteCount`

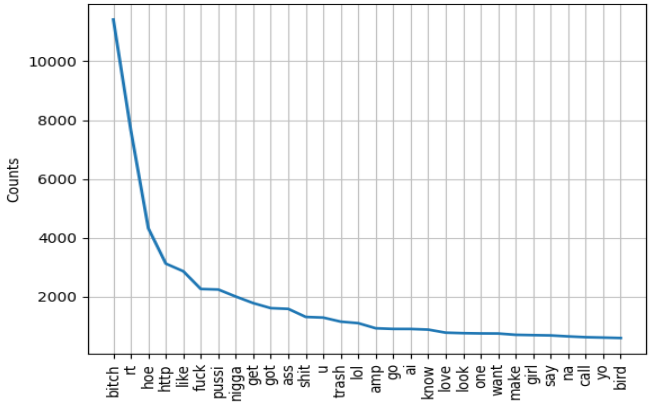
Class Representation Imbalances



Class Representation Imbalances

- Solved by Random Under-Sampling
- Tested all algorithms to both the original and the balanced dataset

Word Frequency

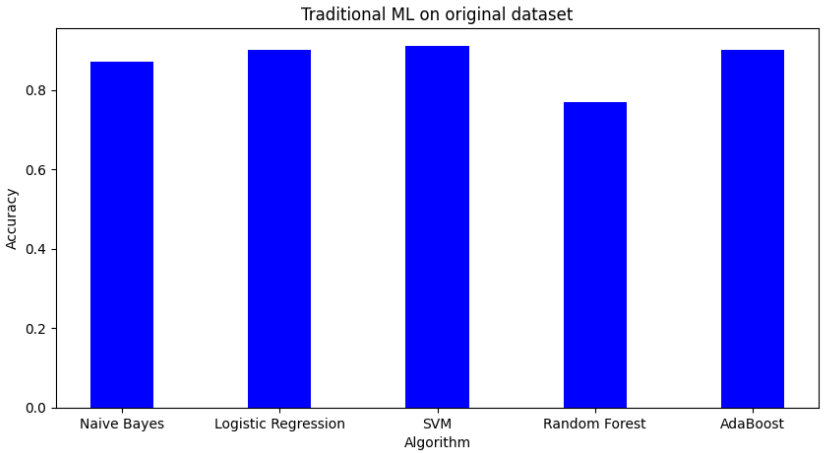


Algorithms Used

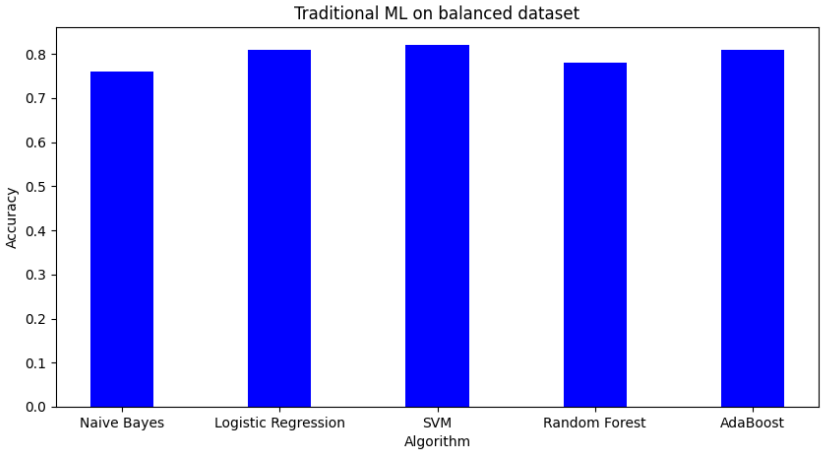
- Traditional ML: Naive Bayes, Logistic Regression, SVM, Random Forest, AdaBoost
- LSTMs
- Using Bert

Results

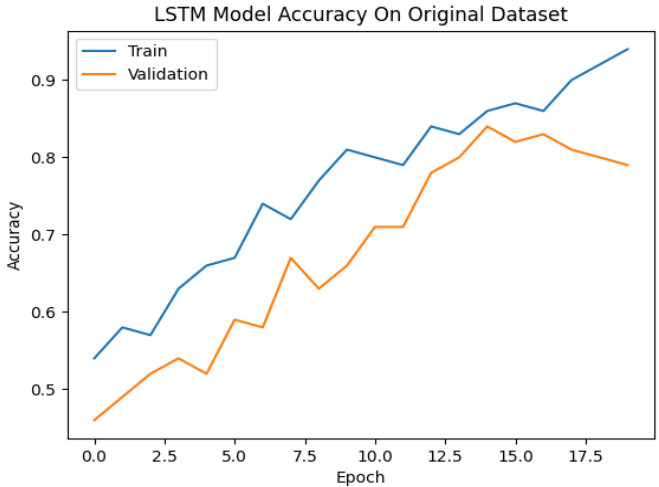
Traditional ML - Original Dataset



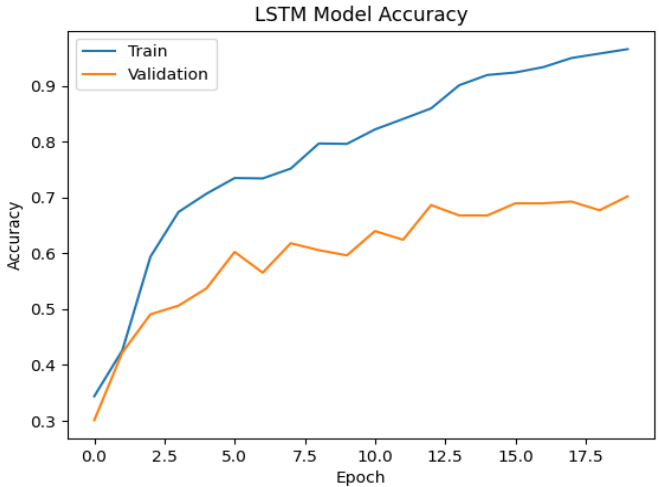
Traditional ML - Balanced Dataset



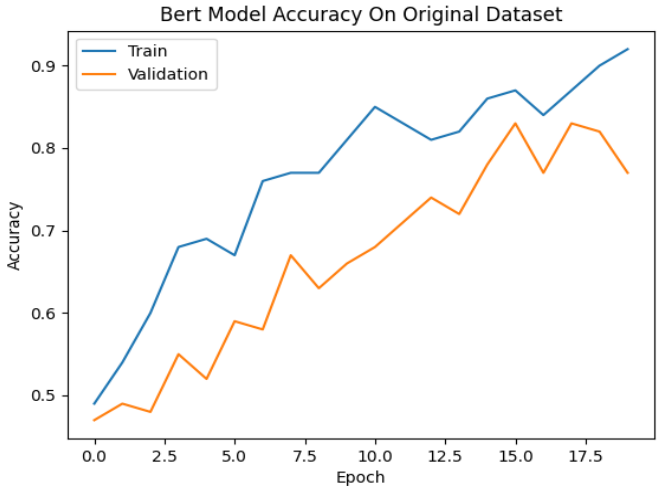
LSTM - Original Dataset



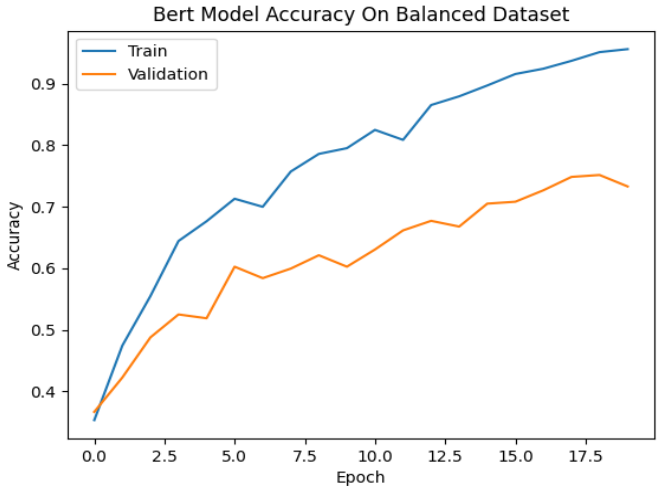
LSTM - Balanced Dataset



Bert - Original Dataset



Bert - Balanced Dataset



Conclusions

Conclusions

- Larger unbalanced Dataset vs Smaller Balanced Dataset Trade Off
- Traditional ML vs LSTMs
- Introduction of Pretrained Transformer to Improve Performance