Hate Speech and Offensive Language Recognition Natural Language Processing

Dimitris Patiniotis Spyropoulos,

July 1, 2022

Table of Contents

- 1 Introduction
- 2 Methodology
- 3 Results
- **4** Conclusions

Hate Speech and Offensive Language Recognition

Introduction

Methodology

Results

Hate Speech and Offensive Language Recognition

Introduction

iviethodology

Results

Conclusions

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

Hate Speech and Offensive Language Recognition

Introduction

Methodology

Results

Recognition

Hate Speech and Offensive Language

meroduction

Methodology

Results

Conclusions

Conclusions

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
```

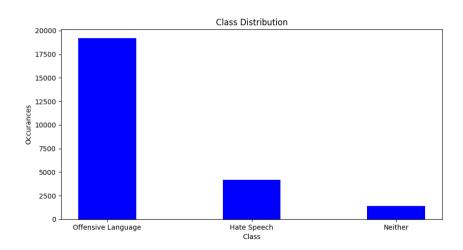
Hate Speech and Offensive Language Recognition

Introduction

Methodolo

Results

Class Representation Imbalances



Hate Speech and Offensive Language Recognition

Introduction

Wiethodolog

Results

Class Representation Imbalances

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

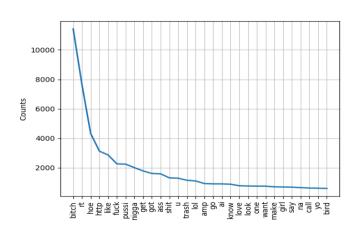
Hate Speech and Offensive Language Recognition

Introduction

Methodology

Results

Word Frequency



Hate Speech and Offensive Language Recognition

Introduction

Methodolog

Results

Algorithms Used

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

Hate Speech and Offensive Language Recognition

Introduction

Methodolog

Results

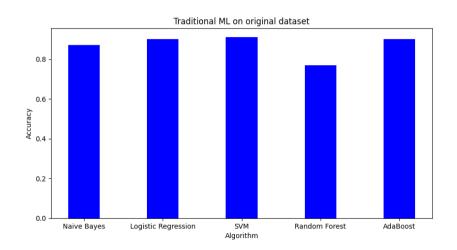
Hate Speech and Offensive Language Recognition

introduction

. .

Results

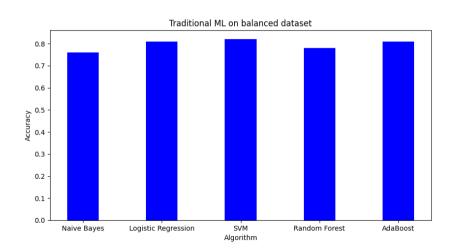
Traditional ML - Original Dataset



Hate Speech and Offensive Language Recognition

ntroduction

Traditional ML - Balanced Dataset

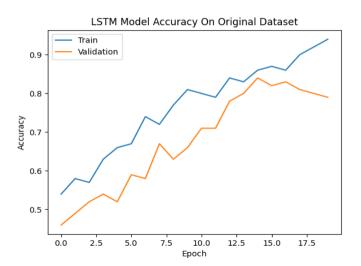


Hate Speech and Offensive Language Recognition

itroductioi

Results

LSTM - Original Dataset

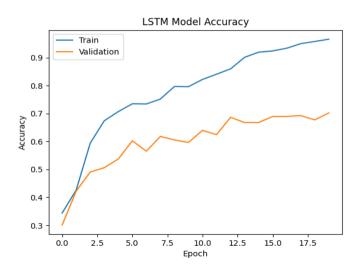


Hate Speech and Offensive Language Recognition

Methodolo_l

Results

LSTM - Balanced Dataset

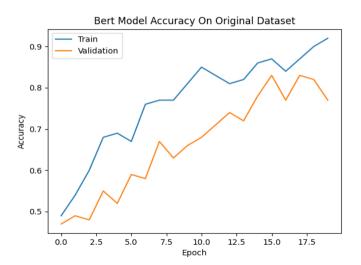


Hate Speech and Offensive Language Recognition

Methodolo

Results

Bert - Original Dataset

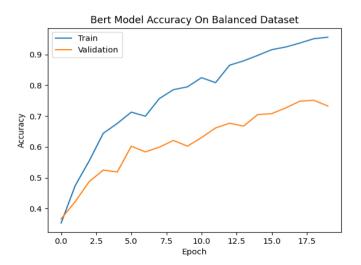


Hate Speech and Offensive Language Recognition

Methodolo

Results

Bert - Balanced Dataset



Hate Speech and Offensive Language Recognition

/lethodolog

Results

Conclusions

Hate Speech and Offensive Language Recognition

Introduction

Methodology

Results

Conclusions

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

Hate Speech and Offensive Language Recognition

Introduction

Methodology

Results