Cross-Lingual Hate Speech Detection

Samrudhi Keluskar
Liverpool John Moores University – Master of Science in Data Science

Agenda

Introduction

Literature Review

Research Methodology

Results

Conclusions

Introduction

What is Hate Speech?

Aim and Objectives of this Study

Significance of the Study

Scope of the Study

Literature Review

Data Sources

- Twitter
- Far-right social networks
- Clubhouse
- Youtube comments

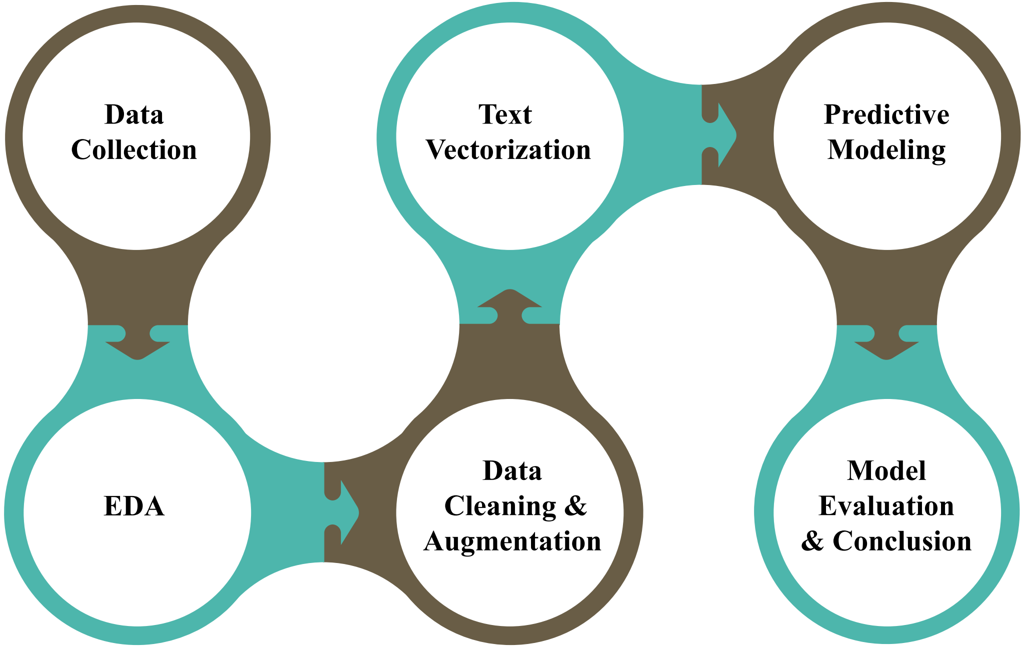
Text Vectorization Methods

- Bag of Words & TF-IDF
- Word2Vec & GloVe
- FastText & LASER
- BERT like LLMs

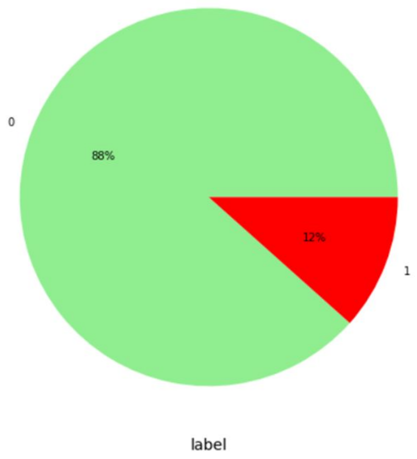
Machine Learning Models

- Logistic Regression
- Random Forests
- Support Vector Machines
- Deep Neural Networks

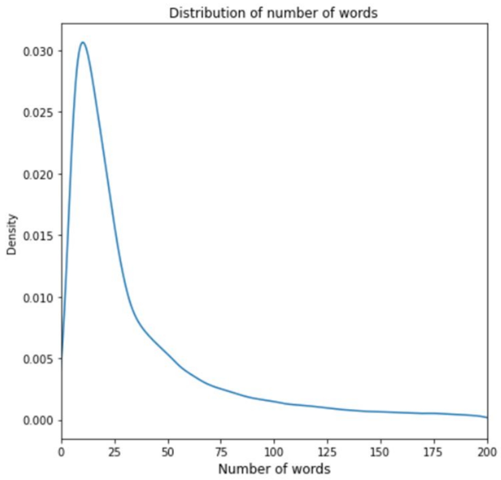
Research Methodology



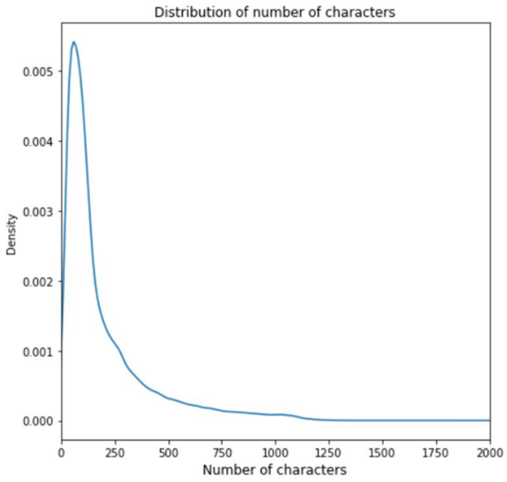
Exploratory Data Analysis



Label Distribution



Distribution of Number of Words

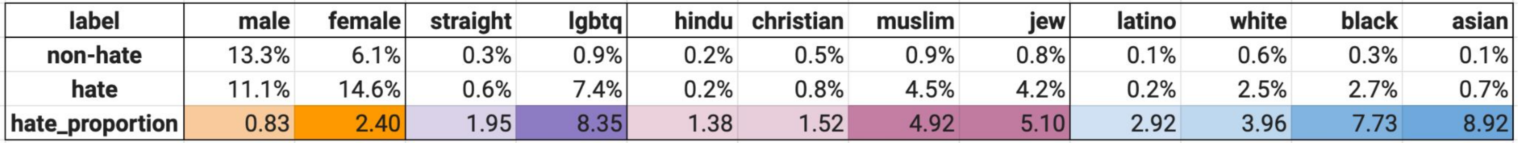


Distribution of Number of Characters

Exploratory Data Analysis

The table below compares the mentions of specific groups in hate vs non-hate sentences.

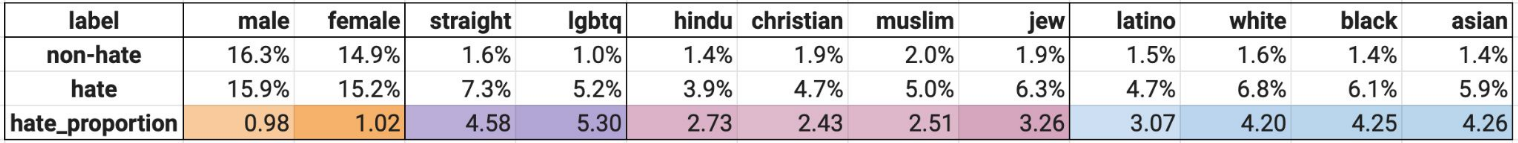
Hate proportion before adding counterfactuals



label	male	female	straight	lgbtq	hindu	christian	muslim	jew	latino	white	black	asian
non-hate	13.3%	6.1%	0.3%	0.9%	0.2%	0.5%	0.9%	0.8%	0.1%	0.6%	0.3%	0.1%
hate	11.1%	14.6%	0.6%	7.4%	0.2%	0.8%	4.5%	4.2%	0.2%	2.5%	2.7%	0.7%
hate_proportion	0.83	2.40	1.95	8.35	1.38	1.52	4.92	5.10	2.92	3.96	7.73	8.92

Table 1

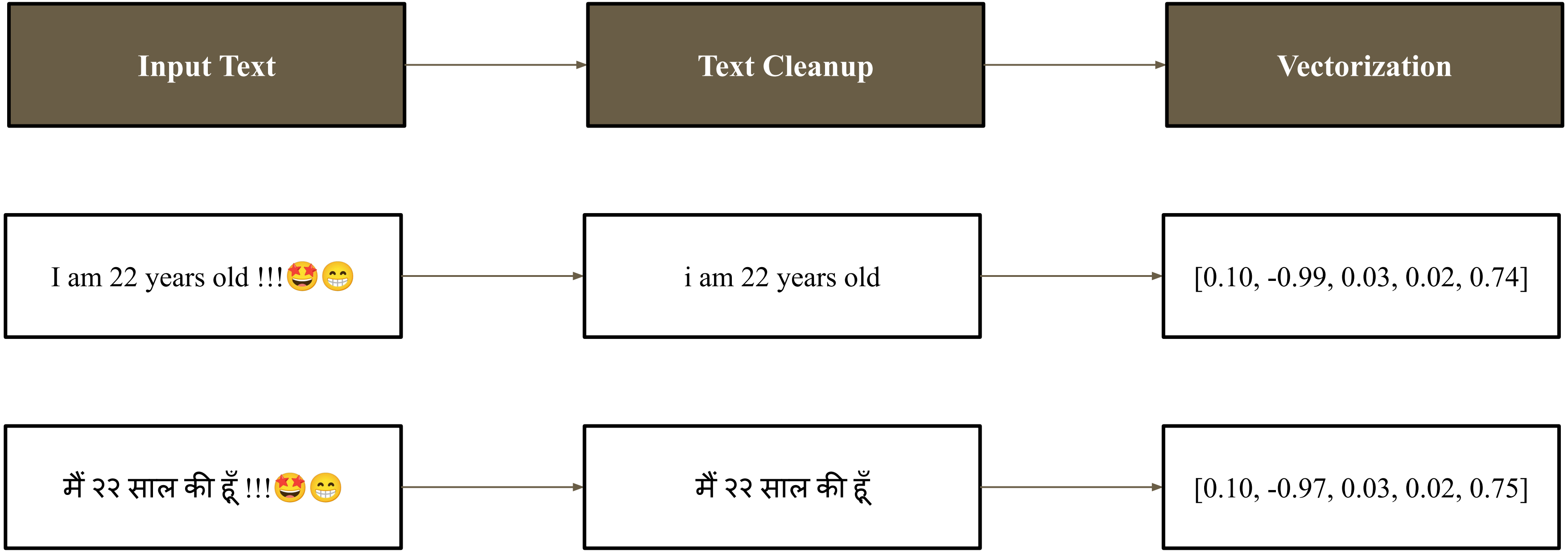
Hate proportion after adding counterfactuals



label	male	female	straight	lgbtq	hindu	christian	muslim	jew	latino	white	black	asian
non-hate	16.3%	14.9%	1.6%	1.0%	1.4%	1.9%	2.0%	1.9%	1.5%	1.6%	1.4%	1.4%
hate	15.9%	15.2%	7.3%	5.2%	3.9%	4.7%	5.0%	6.3%	4.7%	6.8%	6.1%	5.9%
hate_proportion	0.98	1.02	4.58	5.30	2.73	2.43	2.51	3.26	3.07	4.20	4.25	4.26

Table 2

Data Preprocessing



Input Text	Text Cleanup	Vectorization
I am 22 years old !!!🤩😁	i am 22 years old	[0.10, -0.99, 0.03, 0.02, 0.74]
मैं २२ साल की हूँ !!!🤩😁	मैं २२ साल की हूँ	[0.10, -0.97, 0.03, 0.02, 0.75]

Predictive Modeling

Model	Train Set (F1 Score)	Test Set (F1 Score)
Logistic Regression	0.78	0.77
Random Forest	0.98	0.75
Gradient Boosting	1.00	0.87
K-Nearest Neighbors	1.00	0.87
Support Vector Machines	1.00	0.95
Neural Networks (MLP Classifier)	0.98	0.92

Results

Data Language	F1 Score
English	0.99
Portuguese	0.99
Arabic	0.96
Hindi	0.96
Greek	0.93
Chinese	0.92
Danish	0.91
Turkish	0.91

Conclusions

Summary

Contributions

Recommendations

Limitations

Future Work

Thank You!