Hate Speech Detection

NLP Project for Data Glacier by Sarah Littell

Intro:

Group Name: Hate Speech Detection

Name	Email	Country	College/Company	Specialization
Sarah Littell	sarahlit@me.com	USA	Tulane University	NLP

Problem description:

Hate speech is any type of verbal, written and/or behavioral communication that attacks, uses derogatory/discriminatory language against a person or group based on inherent characteristics. For example: based on their religion, ethnicity, nationality, race, color, ancestry, sex or any other identity. In this problem, we will implement a hate speech detection model in order to optimize the speed at which hate speech is identified and removed from a social media platform.

Hate Speech Detection is generally a task of sentiment classification. A model that can classify hate speech from non-hate speech can be achieved by training it on data that is generally used to classify sentiments. For this task of hate speech detection, we will use pre-labeled tweets to train a ML model to identify tweets containing hate speech.

Business understanding:

Hate speech creates an extremely negative environment for users, making them less likely to use the platform and more likely to turn to platforms where they do not experience harassment. This can lead to a significant decrease in the number of users a platform has, and a subsequent decrease in profit and market share, making this an essential problem to get under control for business reasons in addition to moral reasons.

Project life cycle:

Phase	Deadline	Tasks
Data ingestion	2/26/2023	Process data, dedup and validate
Data cleansing	3/2/2023	Data cleansing and featurization (≥2 techniques)
Exploratory Data Analysis	3/9/2023	Get preliminary model ideas, finalize approach plans

EDA Presentation	3/16/2023	Ppt for business users, technical details on the model(s) nailed down
Model Selection/building	3/23/2023	Implement model
Final	3/30/2023	Finish everything

Data Intake report:

Name: NLP Project, Hate Speech Detection

Report date: 02/19/2023

Internship Batch: LISUM17

Version: 1.0

Data intake by: Sarah Littell

Data intake reviewer: Sarah Littell

Data storage location: https://github.com/sarahlittell/NLP_project

Tabular data details:

Total number of observations	31962
Total number of files	1
Total number of features	3
Base format of the file	.csv
Size of the data	3.1 MB

Proposed Approach: Make sure there are no duplicate IDs, make sure if the same tweet exists more than once it is labeled the same both times, then delete all duplicates.

Data understanding:

Type of data: the data comes in a csv file with 3 columns; ID, label, and the tweet itself

Problems in data: the tweet can feature useless information when determining if it's hate speech such as uppercase letters, punctuation, and special characters

Approaches to overcome issues: I will remove all punctuation and special characters from all tweets, then once there are only letters I will convert the tweets to all lowercase to avoid any false negatives because of capitalization.

Data cleansing/transformation:

Type of data: the data comes in a csv file with 3 columns; ID, label, and the tweet itself

Problems in data: the tweet can feature useless information when determining if it's hate speech such as uppercase letters, punctuation, and special characters

I cleaned and transformed the data by making all the tweets lowercase, removing usernames, punctuation, and special characters. I transformed the data through tokenization, removing stop words, and lemmatization. I then performed feature extraction using Tfidvectorization.

EDA performed on the data:

Quantitative analysis:

Num pos/hate speech tweets vs not

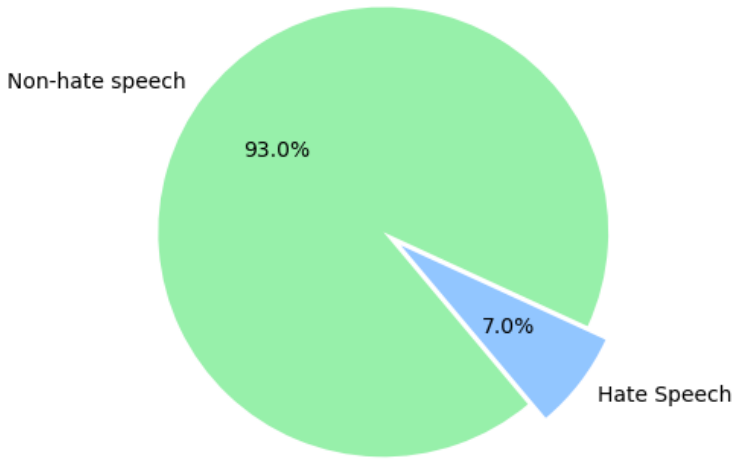
Qualitative analysis:

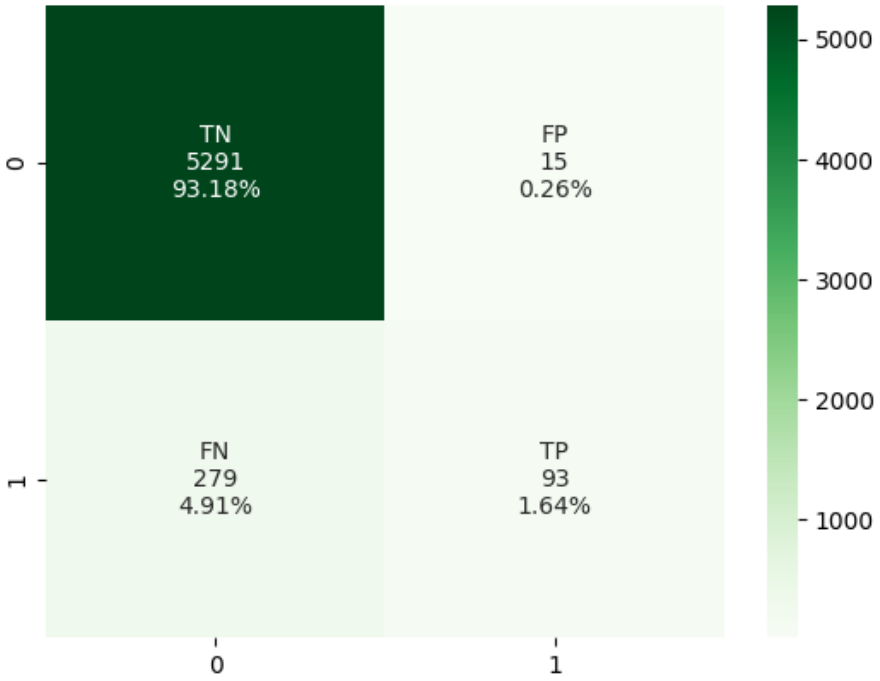
Hate vs non hate speech after cleaning

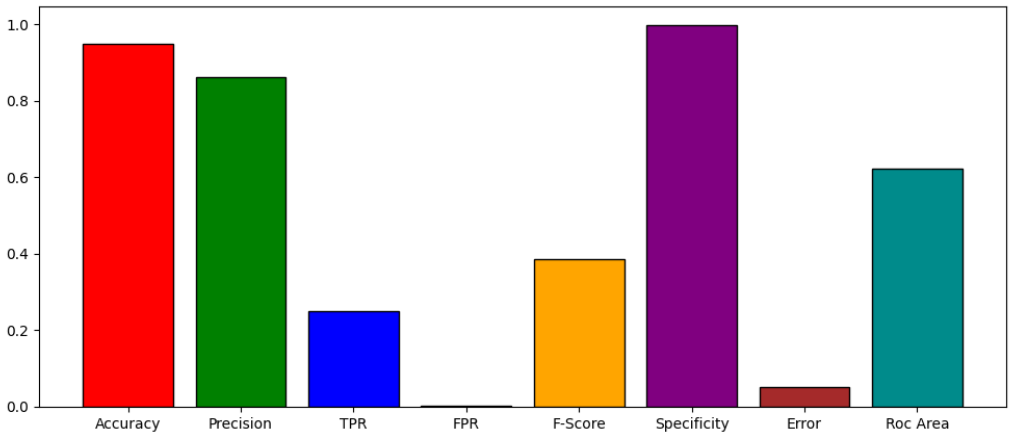
EDA Results:

Non-Hate Comments

Hate Comments


Non-hate speech
93.0%
7.0%
Hate Speech


TN
5291
93.18%
FP
15
0.26%
FN
279
4.91%
TP
93
1.64%
0
1
0
1
5000
4000
3000
2000
1000


1.0
0.8
0.6
0.4
0.2
0.0
Accuracy
Precision
TPR
FPR
F-Score
Specificity
Error
Roc Area

Initial Recommendation:

I plan to exclusively use features identified as significant/impactful using the above methods. Following this, I will build a cnn lstm model according to previous plans with changes along the way to ensure the best performance.

Final Recommendation:

Upon attempting to build the cnn lstm model significant issues were encountered, so a logistic regression model was implemented instead in order to have an adequate model in time.