Inappropriate Speech Detection

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Asiyah Ahmad, Daqi (Jen) Chen, Sam Ferguson
Comp 590: Introduction to NLP

Overview

We are attempting to develop a language model that can identify problematic speech patterns. This issue is important for large social media platforms because their users could be exposed to certain content they'd rather not see.

Solving this problem could help:

- Protect users from targeted harassment.
- Shield younger audiences from inappropriate user content.
- Enhance overall user experience while browsing the platform.

Why is it Important?

Results of our analysis could be applicable to any social media platform and help protect vulnerable users against cyberbullying, etc.

Models

We used a few modeling methods learned in class:

- **Naive Bayes**
- **Logistic Regression**
- **BERT,** state-of-the-art language model

To further improve training, we applied **logistic regression** with an L2 penalty to select the best features, then trained a **Linear Support Vector Classification (SVC)** model, which optimizes the Hinge Loss, with those features to predict the labels.

Conclusion

BERT had the best precision of >98%, meaning it generally doesn't falsely classify safe tweets as inappropriate.

Naive Bayes had the best recall score of >98%, but precision is quite low, meaning it is really conservative, to the point of misclassifying many safe tweets to be inappropriate.

Linear SVC didn't exceed the others in performance, but it reduced input dimensions and hence making the model building less complex while maintaining an accuracy close to LogReg and BERT.

At the end, we obtained a best accuracy of **95.7%** with **Logistic Regression**.

Project & Data Overview

Standards of acceptable speech can vary dramatically across different communities and social settings. Even within a given group, individuals will have different opinions/preferences regarding behaviors surrounding speech.

Our dataset is a collection of tweets tagged by multiple human raters via *crowdsourcing*. They were categorized into the following: "Hate Speech", "Offensive", and "Clean". We combined "Hate Speech" and "Offensive" as "**Inappropriate**", the rest are considered "**Safe**" in our analysis.

Tweet Preprocessing

We used the following methods to preprocess our data:

- Train (*17348*) / Test (*7435*) Split
- Stemming
- **URL** & **User Mention** Tagging*
- Keeping Stopwords
- Keeping Punctuations

RT **@VUULibrary**: It is Shakespeare's birthday. Take a look at some Shakespearean works. http://t.co/JSy3UFfaIb

RAW

rt **usertag** : it is shakespear 's birthday . take a look at some shakespearean work . **urltag**

PROCESSED

Data Encoding

After re-labeling, some examples (text, response class) given below:

"@BabyAnimalPics: baby monkey bathtime http://t.co/7KPWAdLF0R"

0 - SAFE

"@DionaIrish: I hate a ""I'm pregnant"" type of b****."

1 - INAPPROPRIATE

BoW Feature Extraction

To choose the best possible vectorial representation of the tweets for running our models, we first used **CountVectorizer**, then **TfidfVectorizer** from `sklearn`.

Results

```
                                        Model Type
Performance Metric  Naive Bayes    Logistic Reg   BERT            Linear SVC
         Precision  0.9251148545   0.9779050736   0.9839164317    0.9756299734
            Recall  0.9806818181   0.9699675324   0.9633116883    0.9553571428
          F1 Score  0.9520882584   0.9739201303   0.9735050447    0.9653871391
  Overall Accuracy  0.9182246133   0.9569603227   0.9565568258    0.9432414256
```

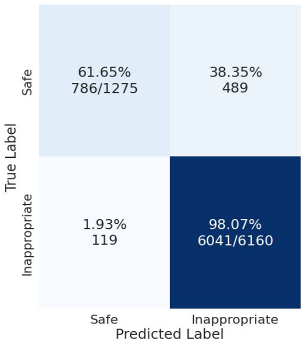

Figure 1. Naive Bayes

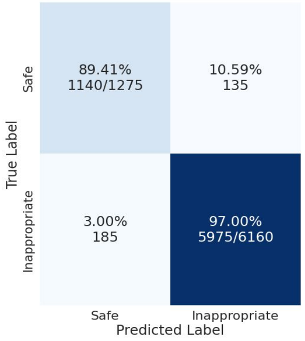

Figure 2. Logistic Regression

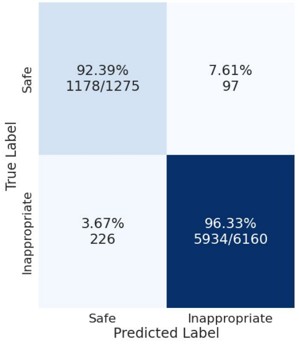

Figure 3. BERT

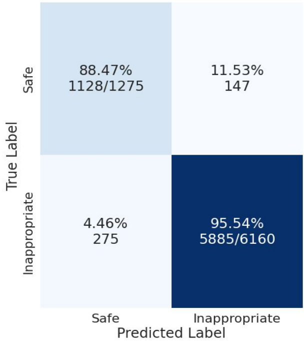

Figure 4. Linear SVC

Demo

Index	Text for Prediction	NB	LogReg	BERT	Linear SVC
0	'.@becky its a good day to be a tarheel'	1	0	0	0
1	'tf is this dook tenting tradition???'	1	1	0	0
2	'jesus what is this dook tenting tradition???'	1	0	1	0
3	'Some weights of the model checkpoint at bert-base-cased were not used.'	0	0	0	0

View Project on Colab

**URL & User Mention Tagging:* Added procedure for the input of Linear SVC Model.