Fall 2018 ITCS 4111/5111
Introduction to Natural Language Processing
Final Presentation

# Hate Speech Detection

Lavina A Sabhnani
(801036525)

Venkataramana Hegde
(801053604)

# HATE SPEECH DETECTION

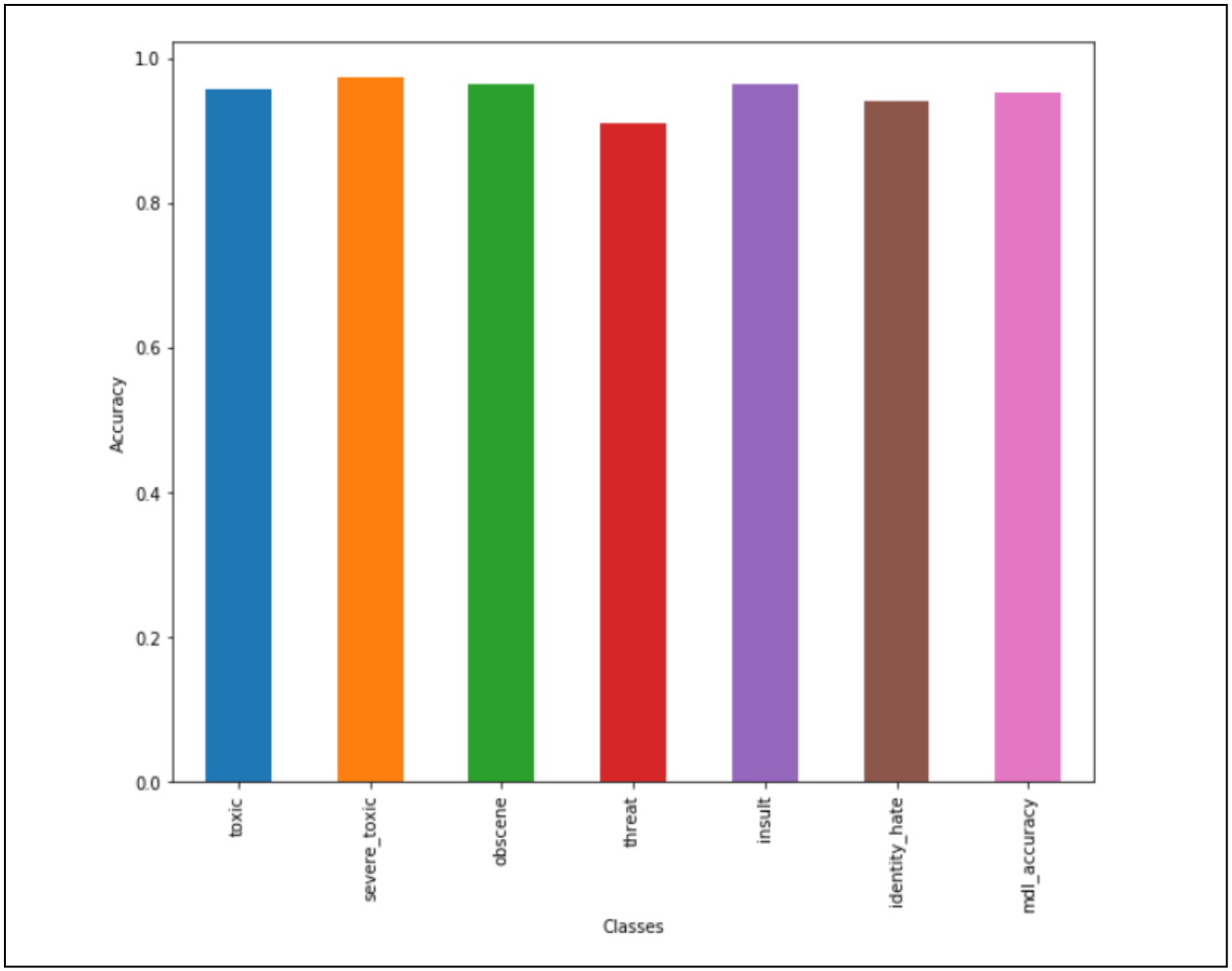
- **Problem Statement:** In United States, hate speech except for obscenity, fighting words, incitement is protected by the First Amendment of the U. S. Constitution assuring the right to free speech, and the internet exercises this right in many formats like blogs, social interactive sites like Facebook, Twitter, Yahoo. In this project we have implemented a model which detects or identifies the hate speech in online text as words submissions are filtered for a fixed list of offensive words, an automatic classifier currently existing isn't publicly available.
- **Dataset:** Our dataset is a Kaggel toxic comment classification dataset with train dataset with 159571 comments and test dataset with 153164 comments. The train.csv constitutes of following attributes: id, comment_text, toxic, sever_toxic, obscene, threat, insult, and identity_hate. The test.csv comprises of the same attributes as that of the train dataset.
- **Testing Dataset:** The Twitter Train dataset comprises of 17348 comments and Test dataset comprises of 7434 comments. The train.csv and test.csv constitutes of following attributes: id, comment_text, hate speech, offensive language and neither.

# Project Implementation

- We've created a model by implementing Logistic regression along with Naïve Bayes function as classifier to implement the following features: NGrams, TFIDF and Word Embeddings.
- The project is implemented in four steps: Loading the data, Cleaning the data, Feature Implementation, Build and Train Model, and lastly Validating the model.
- We compare our model with Naïve Bayes classifier for the features implemented to compare the models for better accuracy and also tested the model on a different Twitter dataset.

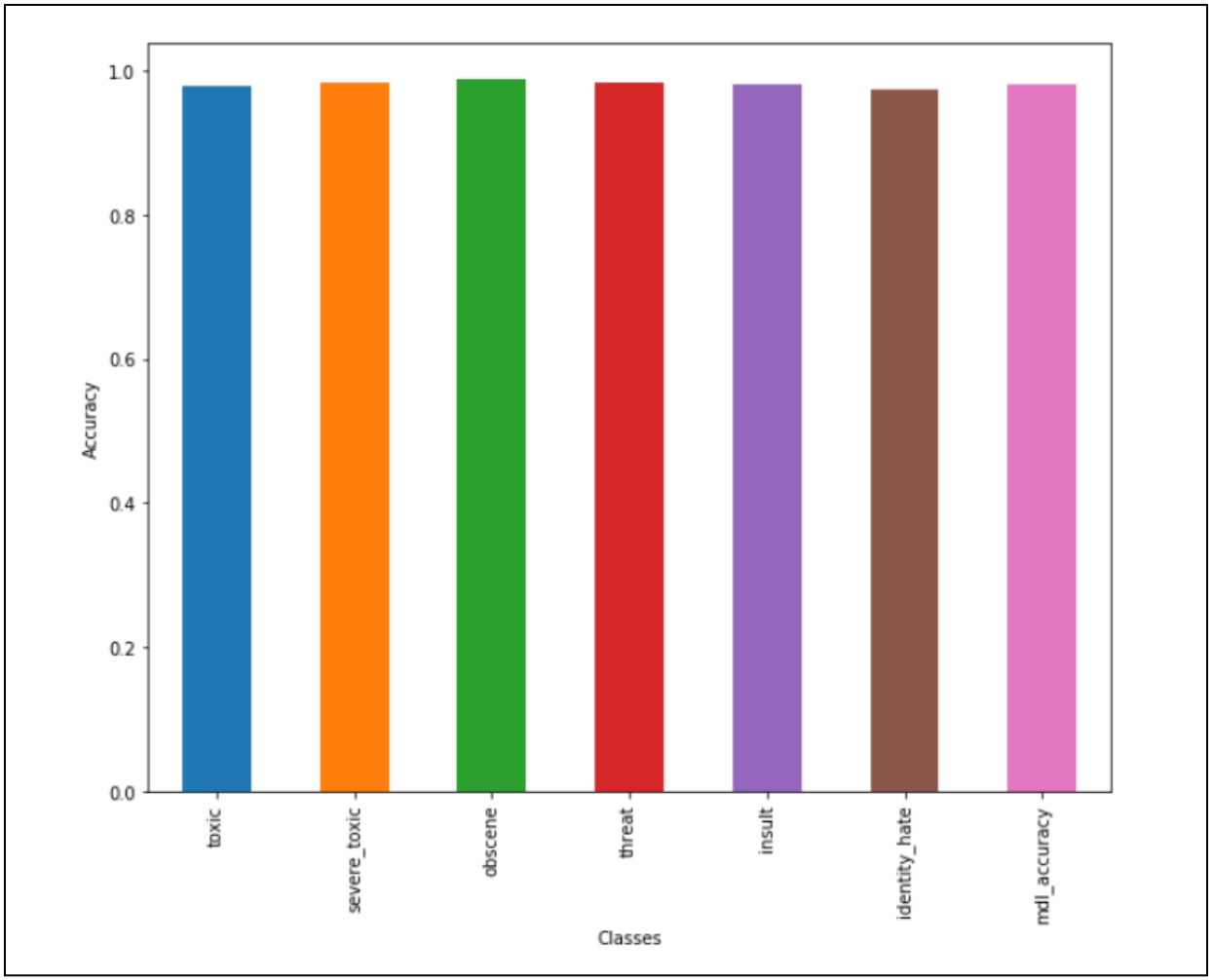
# TFIDF with Ngrams for Naïve Bayes Model.

• For TFIDF with NGram given as parameter the CV score using Naive Bayes classifier obtained is 95.150%



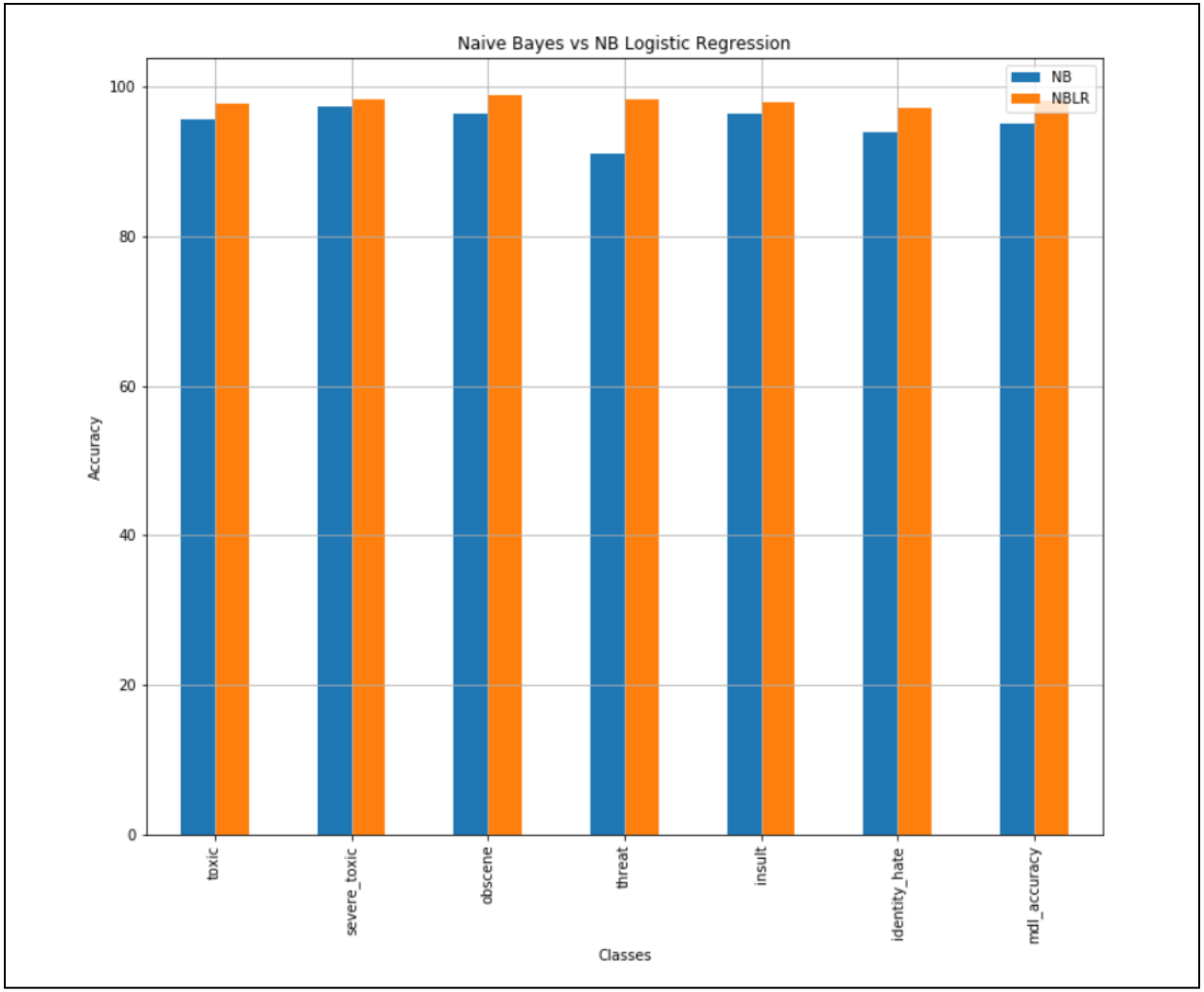
# TFIDF with Ngrams for Logistic Regression- Naïve Bayes Model.

•For TFIDF with NGram as parameter, the model that we built- Logistic Regression with Naïve Bayes (NB-LogReg), the CV score obtained is 98.13%.



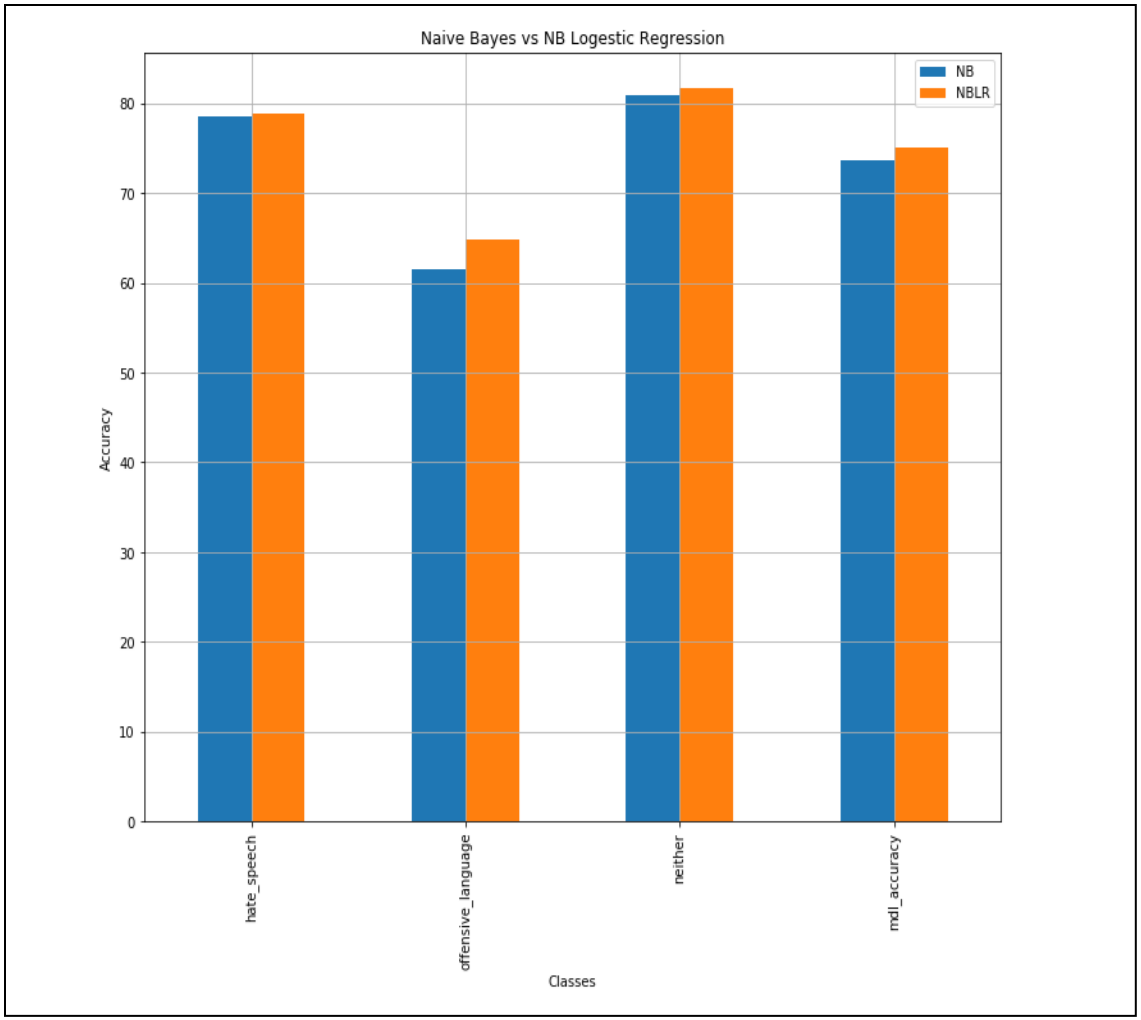
# Comparing both the models CV scores:

- The accuracy difference is around 2.98%



# For Twitter Dataset:

•The accuracy difference is around 1.51% , accuracy for Naïve Bayes model is 73.60% and for Logistic Regression with Naïve Bayes is 75.11%



# Learning's and Future Work

- Modeling of Classifiers
- Functioning's of different types of Ngrams with respect to datasets
- Validation Algorithms
- Different Feature implementations
- Future work: Building of an interactive application based on the model such that it'll take online comment text and returns one class indicating the type of label it categorizes under like toxic or insult.