# Classifying Political Affiliation of Twitter Users using Sentiment Analysis and Topic Models


- Saurabh Katkar
- Mitali Nandargikar
- Akshaya Hebbar

# Problem Statement

- Use of sentiment analysis and topic modeling using search term "**Ebola**" to analyze the level of concern by examining the tweets, and to determine the political affiliation of the user.
- **Hypothesis**
  - Users view about Ebola influenced by media and political personalities.
  - Twitter used to depict sentiment and concern.

# Our Approach

Consider a list of 15 different republican/democratic personalities on Twitter

↓

Extract followers of these personalities

↓

Find a subset of common followers that follow atleast 5 of these twitter users

↓

From this subset extract the tweets that contain the term 'ebola' in them

↓

Once extracted, perform sentiment analysis to find the score of the extracted tweets

↓

Extract different attributes such as gender, word count, frequency of certain terms etc

↓

Store the extracted data in a .csv file

# Fetching Twitter Data

- **Python Twitter API** is a Python wrapper around the Twitter API and the Twitter data model.
- Provides **RESTful** API methods.
- **OAuth 2** authentication used to make read-only calls to Twitter.
- Twitter data collected from mid-October till early November, when Ebola outbreak was quite prevalent in the media.

# Harvesting Common Followers

- Common followers of Republican and Democratic personalities harvested for implicit labelling.
- **GET followers/ids** method of the Twitter API used which returns a cursored collection of user IDs.
- Followers common to 7 out of 15 republican/ democratic personalities were considered.

# Harvesting the Target Tweets

- **GET statuses/user_timeline** method of the Twitter API used to return a collection of the most recent Tweets posted by the user indicated by the **screen_name** or **user_id** parameters.
- Tweets fetched were filtered using **Regular Expression** (import re) methods: **compile** and **search**.
- Thus this methodology was used to harvest the tweets from common users using the search term **'Ebola'**.
- Other terms that divides political ideologies also considered.

# Computing the Sentiment Score

- AFINN wordlist used: scores of 2477 terms
- Tweets containing Ebola:
  - Average score of -0.523 amongst republicans
  - Average score of 0.461 amongst democrats
- Other topic models considered:
  - Issues in favor resulted in a positive average sentiment score and opposed issues resulted in a negative average sentiment score.

# Computing other attributes

- **Gender:**
  - **Using GET users/lookup** method, user name derived from User ID
  - Gender information extrapolated using **United States Census Data**

- **Other attributes derived:**
  - Number of words in a tweet.
  - Frequency of certain words in a tweet. Words considered that displays ‘perspective bias’

# Dataset Derived

- Two datasets derived:
- Only using ‘Ebola’ as a means for deriving sentiment score.
  - Dataset instances: 2000
  - Dataset attributes: 8
- Using ‘Ebola’ along with other terms for deriving sentiment score.
  - Dataset instances: 2000
  - Dataset attributes: 18

# Predicting Labels using Generative and Discriminative Classifiers

- Generative and Discriminative learning :
  - Use different technique in solving classification problem
- Generative classifiers model the joint distribution and learns the model parameters through maximization of the likelihood.
  - Example: **Naïve Bayes Classifier**
- Discriminative classifiers, model the conditional distribution of the class labels given the features, and learn the model parameters through maximizing the conditional likelihood.
  - Example: **Logistic Regression**

# Performance of Algorithms

- Time taken by Logistic Regression Algorithm:

| Logistic Regression Time Elapsed | | | |
|---|---|---|---|
| | user | system | elapsed |
| **Ebola** | 25.41 | 0.08 | 25.64 |

- Time taken by Naïve Bayes Algorithm:

| Naïve Bayes Time Elapsed | | | |
|---|---|---|---|
| | user | system | elapsed |
| **Ebola** | 18.02 | 0.1 | 19.9 |

- **Observation: Naïve Bayes executes faster!!**

# Performance of Algorithm

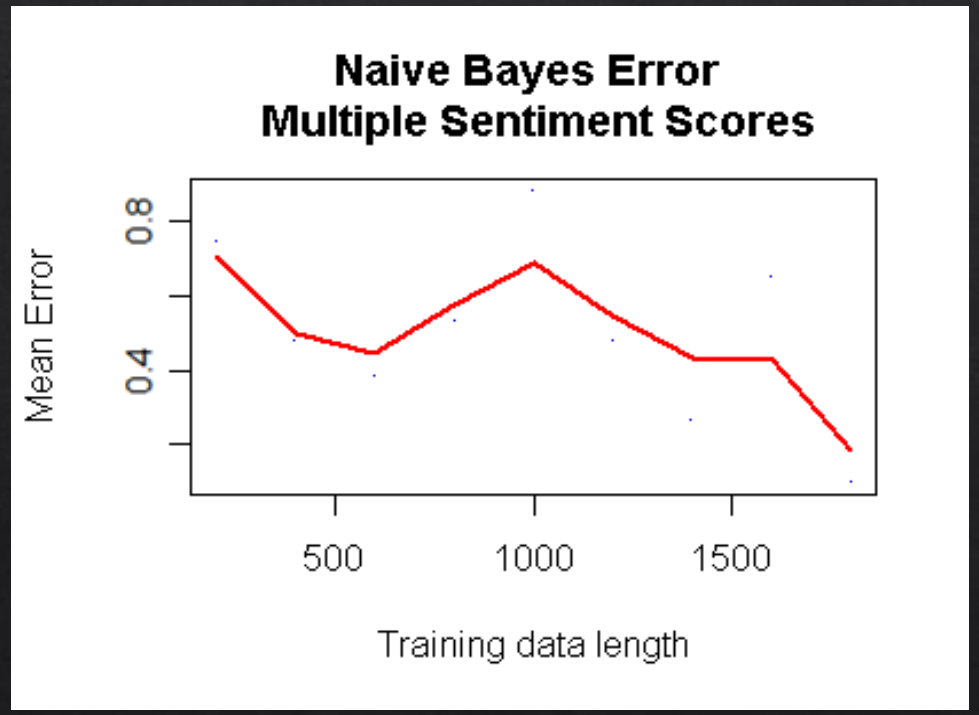
- Performance Measures of Logistic Regression Algorithm

| Logistic Regression Performance Measures | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Dataset** | Accuracy | Precision1 | Precision2 | Recall1 | Recall2 | F1 | F2 |
| **Features: 8** | 0.612 | 0.837 | 0.736 | 0.889 | 0.816 | 0.872 | 0.726 |
| **Features: 18** | 0.814 | 0.833 | 0.901 | 0.935 | 0.855 | 0.820 | 0.961 |

- Performance Measures of Naïve Bayes Algorithm

| Naïve Bayes Performance Measures | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Dataset** | Accuracy | Precision1 | Precision2 | Recall1 | Recall2 | F1 | F2 |
| **Features: 8** | 0.35 | 0.855 | 0.215 | 0.543 | 0.765 | 0.687 | 0.335 |
| **Features: 18** | 0.65 | 0.409 | 0.616 | 0.891 | 0.747 | 0.687 | 0.561 |

# Performance of Algorithm




# Inference

- Naïve Bayes approaches its asymptotic error without the need for a large number of training examples, and it does so very quickly.
- If the number of training examples is very large, then using Logistic regression is comparatively costly, since the parameters of the algorithm are needed to be optimized using gradient optimization
- Logistic regression, on the other hand, is capable of outperforming naive Bayes, given the number of training examples is large enough.

# Conclusion

- Dataset using only sentiment score of search term ‘Ebola’ gives less accuracy as compared to using sentiment scores of other terms along with ‘Ebola’.
- Other features such as gender, frequency of certain words based on lexicon and perspective bias effective in determining the political affiliation .
- When dataset is large, Logistic Regression more accurate than Naïve Bayes.
- However Naïve Bayes is faster, thus reaching its error faster.

# Future Work

- A more vast wordlist other than AFINN or a sentiment analysis API would provide better results for sentiment analysis.
- Many other attributes such as ethnicity, geolocation can be derived and taken into consideration.
- **Hybrid generative-discriminative techniques**
- **Inclusion of prior knowledge in discriminative methods.**

# Thank You

- Questions?