TOPIC CLASSIFICATION

NLP Project

Supervisor:

Dr. Nermin Kamal Negied

Team Members:

1. **Karim Mohamed Ibrahim**
2. **Omar Tarek Merghany**

Introduction:

Text classification is one of the problems that appeared in the natural language processing and big data analysis. It is mainly about how to assign each text to a specific class or category. Some problems are binary classification where the number of class are two such as spam or non-spam mails problem. On the other hand, there are multiple classes' classification such as movie genre classification. By using some NLP and machine learning algorithms, this problem can be solved.

Problem Statement:

Twitter has become as much of a news media as a social network, and much research has turned to analyzing its content for tracking real-world events, from politics to sports and natural disasters. Twitter users tweet their views in the form of short text messages.

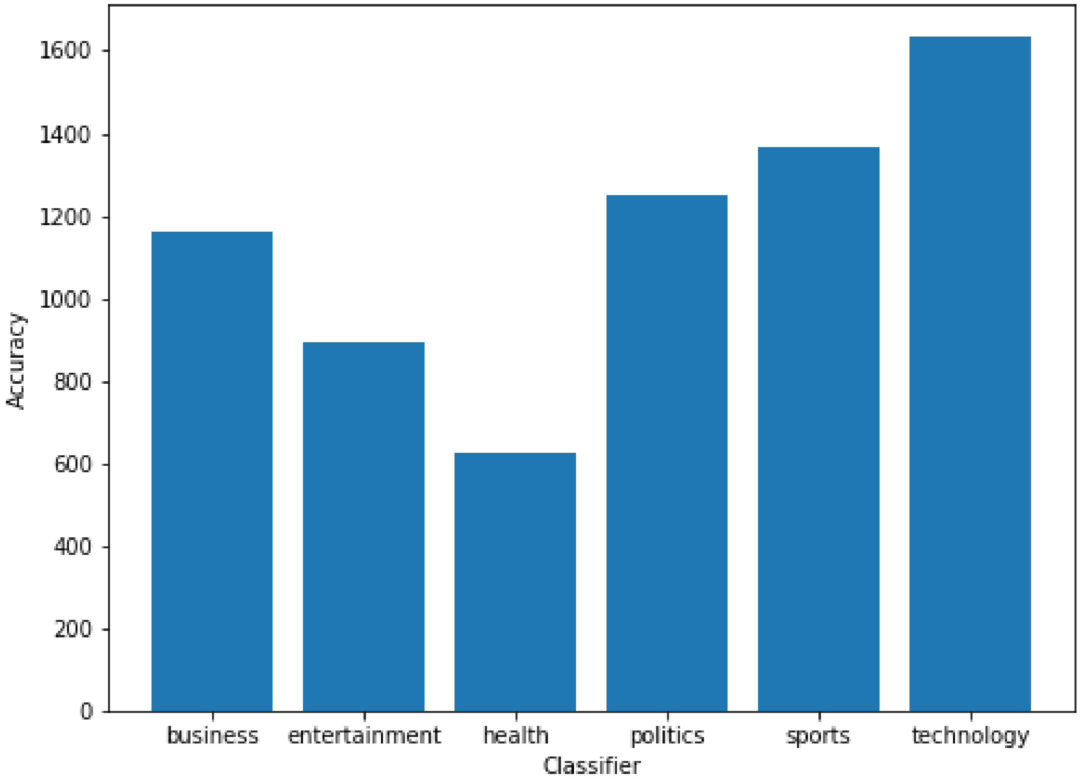
Twitter generates 340+ million tweets per day, twitter is becoming a major source of information. This makes tweet classification a challenging problem for the Machine Learning researchers Twitter topic classification is classifying the tweets in to a set of predefined classes. We have a dataset of tweets and need to assign each tweet to the appropriate tags such as

1. Technology
2. Business
3. Politics
4. Entertainment
5. Sports
6. Health

Dataset:

The Dataset was created by fetching titles of different Subreddit relating to 6 main categories (Business, Entertainment, Health, Politics, Sports, Technology).

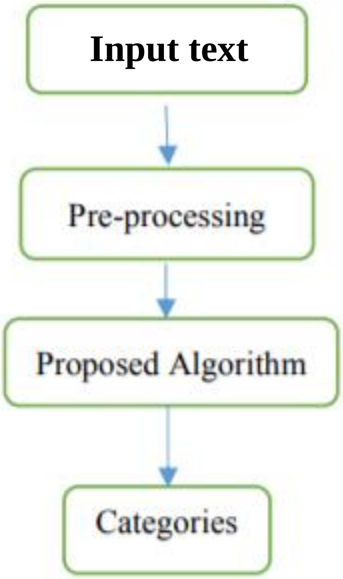
The Dataset was splitted to train and test.



We tested our model on the test set of Subreddit, in addition to test set collected from Twitter.

Problem Solution:

1. Feature extraction using natural language processing algorithms.
2. Use the extracted features to train the Multinomial Naive Bayes Classifier.
3. Use the extracted features to train SVM classifier.
4. Compare between the performance of each classifier



Procedures and Analysis

After we have collected the data from the Reddit, we have to make some preprocessing to the train data (labeled) and the test data (unlabeled) to construct a bag of words from the trained data so that we can evaluate the unlabeled data according to the bag of words collected from the trained tweets.

- Cleansing

In this step, we have removed the URLS, special characters, numbers and stop words. The data should be filtered from stop words before the classification phase, as the stop words do not express any class in our problem. Besides, if they were left and included in the classification stage, they may influence the accuracy of the training. These words are such as:
'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', etc.

- Text processing: Tokenize & Transform to lowercase

The text is tokenized then transformed to lowercase so that it can be used for the test.
After filtering data, features were extracted from the training dataset. Using NLTK library, the data has been tokenized to a set of words. These words are the features that can express which text belongs to which class.

- Build word list for Bag-of-Words

The wordlist (dictionary) was built by simple count of occurrences of every unique word across all of the training dataset. The most common words are the typical English stop-words. We have filtered them out, however

The wordlist is also saved, so the same words can be used for the testing set.

Machine Learning Algorithms

- Naive Bayes Algorithm:

Naive Bayes Algorithm depends on the idea of the conditional probability to determine the class of each tweet.

$$p(c|x) = \frac{p(x|c) * p(c)}{p(x)}$$

To determine which class that text belongs to, we need to calculate the probability of that class given that this tweet. That means we need to calculate the likelihood p(x|c) and the probability of the class p(c). We can ignore dividing by p(x) as it is fixed.

The data was expressed into frequency array that expresses how much time each feature appears in each text. Then, the class probability and likelihood probability can be calculated.

- SVM Algorithm:

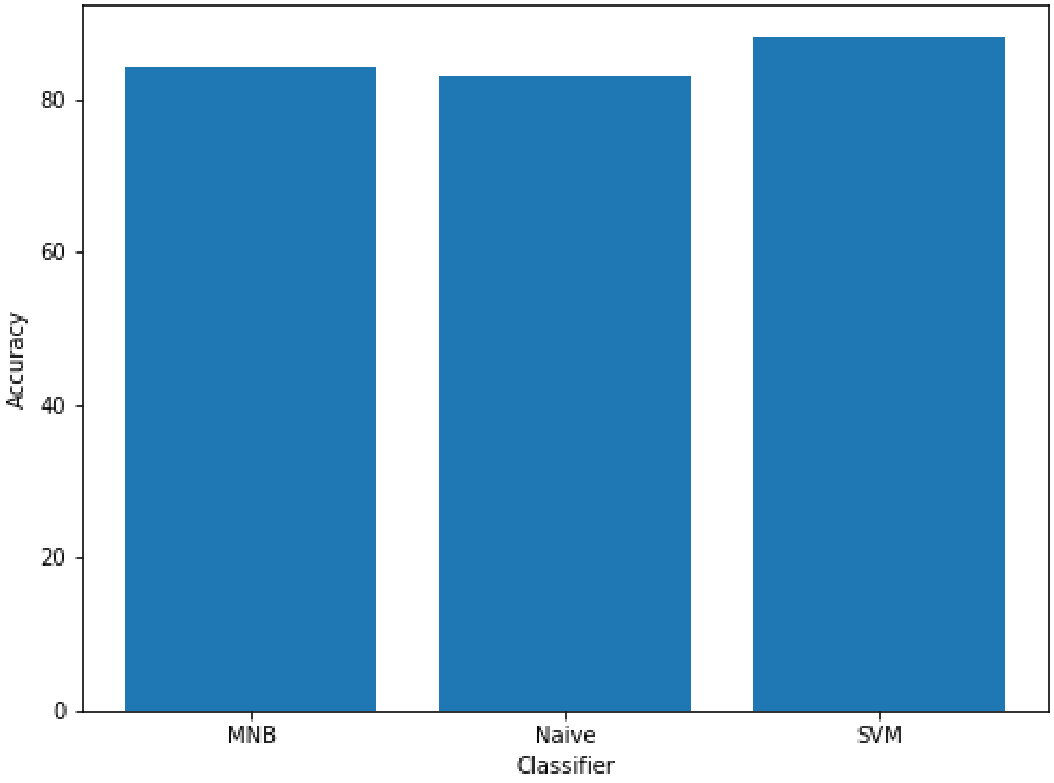
SVM algorithm depends on finds the hyperplanes that separate the classes from each other. To implement this algorithm, the feature vector of each dataset represent a dimension and the SVM algorithm tries to separate them according to their class.

Results

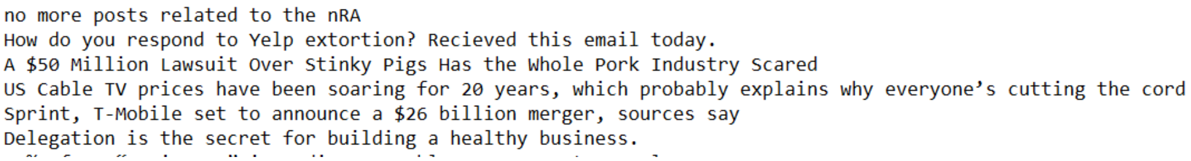
Linear SVM classifier accuracy: 88.0

Naïve Bayes accuracy: 83.0

MNB classifier accuracy: 84.0

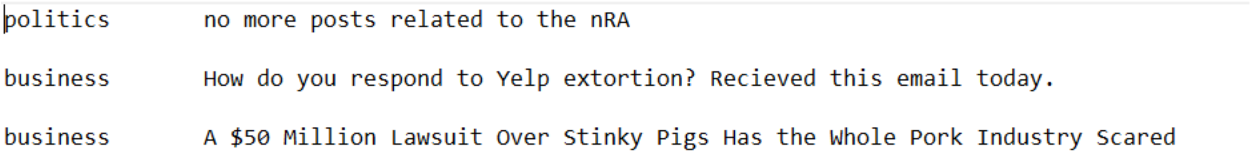


This is a sample from the test set



```
no more posts related to the nRA
How do you respond to Yelp extortion? Recieved this email today.
A $50 Million Lawsuit Over Stinky Pigs Has the Whole Pork Industry Scared
US Cable TV prices have been soaring for 20 years, which probably explains why everyone’s cutting the cord
Sprint, T-Mobile set to announce a $26 billion merger, sources say
Delegation is the secret for building a healthy business.
```

Naïve

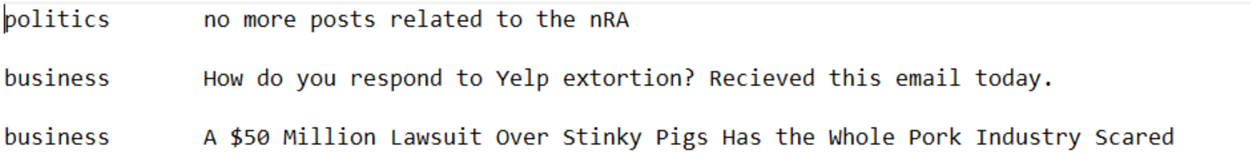


```
politics        no more posts related to the nRA

business        How do you respond to Yelp extortion? Recieved this email today.

business        A $50 Million Lawsuit Over Stinky Pigs Has the Whole Pork Industry Scared
```

MNB

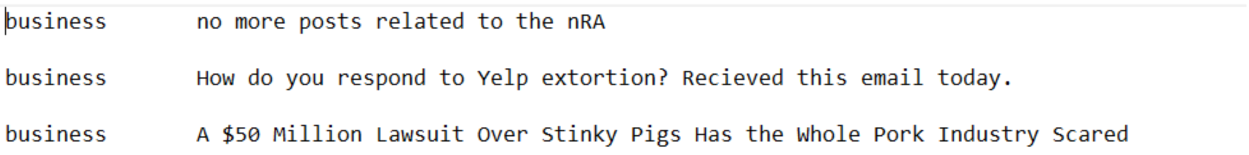


```
politics        no more posts related to the nRA

business        How do you respond to Yelp extortion? Recieved this email today.

business        A $50 Million Lawsuit Over Stinky Pigs Has the Whole Pork Industry Scared
```

SVM



```
business       no more posts related to the nRA

business       How do you respond to Yelp extortion? Recieved this email today.

business       A $50 Million Lawsuit Over Stinky Pigs Has the Whole Pork Industry Scared
```

Discussion

The result was not as expected as the model consider each word (unigram) and does not take into consideration the context and the entire sequence, so when the word “White House” is mentioned in a context of business the text would be classified as “Politics” meanwhile it is a “Business” class. In addition, the trained data should be bigger to be able to consider Neural Nets models.