TWITTER SENTIMNET ANALYSIS

Surya Nikhita Naramsetti & Sravya Busayavalasa
CS-583 Data Mining and Text Mining, Fall-2019
Department of Computer Science
University of Illinois at Chicago, Illinois

ABSTRACT:

Data mining and Text Mining is also called knowledge discovery in databases (KDD). It is commonly defined as the process of discovering useful patterns. Web contains huge volumes of unstructured text data and analyzing such text to find sentiment is called Sentiment Analysis.

Sentiment analysis is an important task in natural language understanding and has a wide range of real-world applications. The typical sentiment analysis focus on predicting the positive neutral or negative polarity of the given sentence(s).

Primary task of this project is to perform sentiment analysis over a corpus of tweets during the U.S. 2012 Re-Election about the candidates Barack Obama and Mitt Romney using different machine learning algorithms. We attempt to classify the polarity of the tweet where it is either positive or negative. If the tweet has both positive and negative elements, the more dominant sentiment should be picked as the final label. For better classification we need to extract useful features from the text such as unigrams.

TECHNIQUES:

Data Preprocessing:

- **Replacing [comma]:** All the occurrences of the regular expression [comma] in the sentence are replaced by the actual character ","
- **Stripping off white spaces:** Additional white spaces are stripped off from each sentence
- **Removal of stop words:** Maintained a separate file of all possible stop words of English language. These do not contribute towards the classification
- **Removal of other special characters:** All special characters that are preceding or succeeding words in the sentence that might hinder with word recognition and creating bag of words are removed. E.g. '' "" ! ? ; % ( ) [ ] $ @ / \ - etc
- **Conversion to lower case:** All the sentences are converted to lower case to maintain uniformity
- **Removal of tweets with irrelevant classes:** All tweets which are unlabeled or having irrelevant class labels are removed.
- **Removal of URLs:** Users often share hyperlinks to their web pages in their tweets. These URLs might not be important for text classification and hence were replaced with spaces.
- **Hashtags and Emoticons:** Contents of hash tags were retained except for #. Emoticons were replaced with spaces.

Feature Extraction:

The features used in the process are the words after applying all the preprocessing steps to the data. These words as the features will be used for training which results in building the classifier. The test set is later fed through the model created by the pipeline to classify the sentences and generate the Accuracy, Precision, Recall and F-score for each class/sentiment. The unigrams of text is considered to be feature which thus results in ignoring the position of sentence. The sentence is instead considered as a "bag of words". Also, in pipeline the other feature that is used is TF-IDF i.e. Term Frequency-Inverse Document Frequency. TFIDF Vectorizer of the library sklearn.feature_extraction.text is used which uses Count Vectorizer and then TFIDF-Vectorizer to store the count of each word in the separate vector. The TFIDF vectorizer is used to convert the sentence into a vector and then fed to the classifier to train the model.

Data Sampling:

Sampling allows any skewed dataset to become balanced as skewed training results in classification of tweets to mostly to majority class in training data. Hence sampling is performed, and the balanced inputs are provided to the classifier learning.

Because the Romney dataset is highly skewed with majority classes as negative, over-sampling is performed on the dataset using SMOTE (Synthetic Minority Over-Sampling Technique). SMOTE is a library of imblearn.over_sampling which synthesize new data samples of minority class by considering the nearest neighbors of the data point and considers a new data sample on the line joining the data point itself and its nearest neighbor. Which results in a dataset of equal distribution of classes.

CLASSIFICATION ALGORITHMS:

Multinomial Naïve Bayes:

Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem. The descriptive attributes/features are assumed to be conditionally independent from each other, which makes a naïve assumption.

Major strengths of naïve Bayes classifier are: handling noisy data since it is averaged out in the estimation of conditional probability, null values are ignored, and irrelevant features are uniformly distributed so they do not have significant influence on the classification result. Weaknesses are mainly attributed to the assumption of complete independence amongst attributes.

Multinomial Naïve Bayes performed well on both datasets with consistent results before and after sampling.

Support Vector Machine:

Support Vector Machines is another popular classification technique. It constructs a hyper plane or set of hyper planes in a high-dimensional space such that the separation is maximum. The hyper plane identifies

certain examples close to the plane, which are called as support vectors. The results of SVM are consistent and equally good as other traditional classification models like Multinominal Naïve Bayes, Logistic Regression.

Decision Tree Classifier:

A Decision Tree is a flowchart-like tree structure, in which each internal node represents a test on an attribute (features) and each branch represents an outcome of the test, and each leaf node represents a class (positive, negative or neutral). The results of decision tree haven't been consistent. Decision Tree classifier resulted in poor accuracies compared to all the other classifiers for both datasets. The reason behind that is the sparse dataset.

Logistic Regression:

Logistic Regression models are feature-based models. The idea behind this model is that one should prefer the most uniform models that satisfy a given constraint. Logistic Regression is the classifier which produced accurate and consistent results and gave best accuracies compared to other traditional classifiers like Multinominal Naïve Bayes and SVM before sampling and after sampling.

Voting Classifier:

It is an ensemble classifier. Soft Voting/Majority Rule classifier for unfitted estimators. All the classifiers employed are parsed in the voting classifier. The voting classifier in case of soft voting determines the classifiers results and averages the evaluation parameters obtained from those best classifiers. In case of hard voting it considers the majority classifiers vote out of all the classifiers to classify the data sample.

The Voting classifier used is an ensemble of Multinominal Naïve Bayes, Logistic Regression, SVM with hard voting. The results of Voting Classifier are the best out of all other imposed classification models.

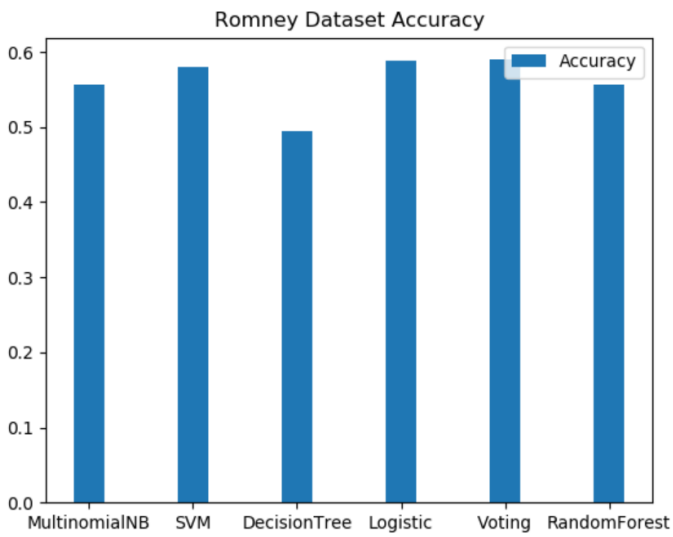
Random Forest Classifier:

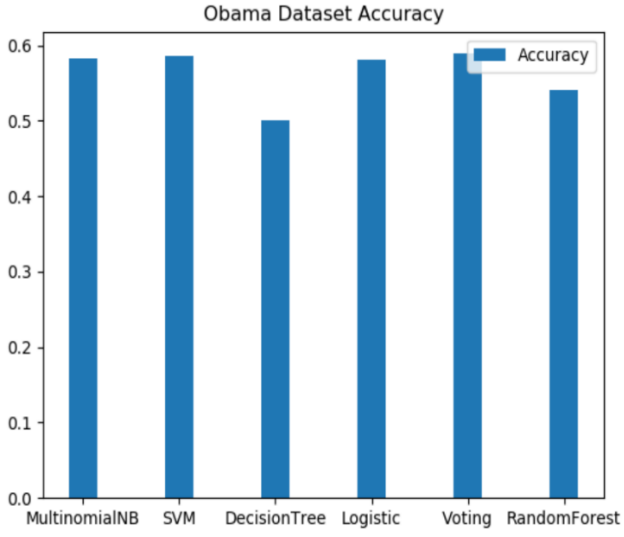
Random forests or random decision forests are an ensemble learning method or classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. The results of Random Forest Classifier are not satisfactory like decision trees but performed better than decision tree on both data sets.

EVALUATION:

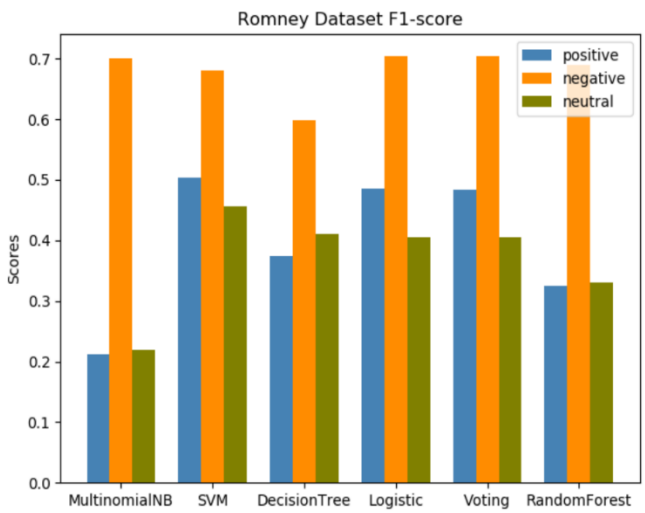
The evaluation parameters of the classifiers are F-Score, accuracy, precision and recall. The below chart shows the Romney and Obama Accuracies for various classifiers.

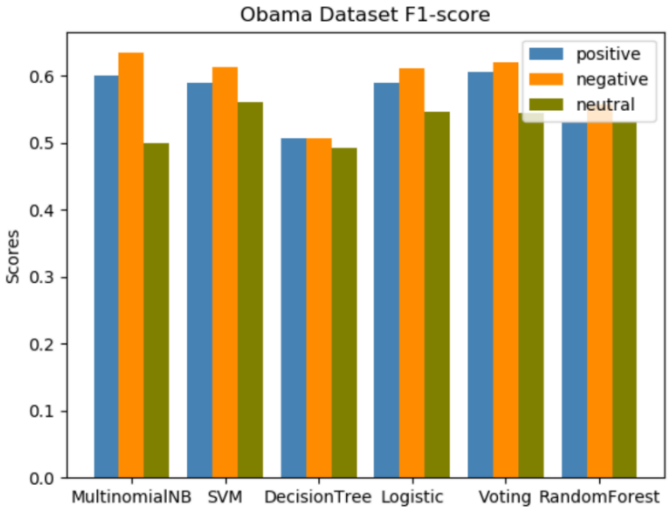
Multinominal SVM, Logistic Regression and Voting classifier are the best classifiers which showed significant results than the rest of the classifiers on the Obama dataset and Romney dataset.



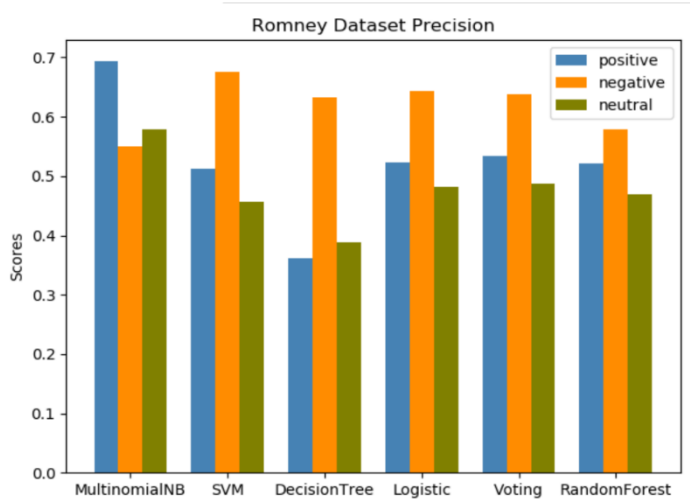


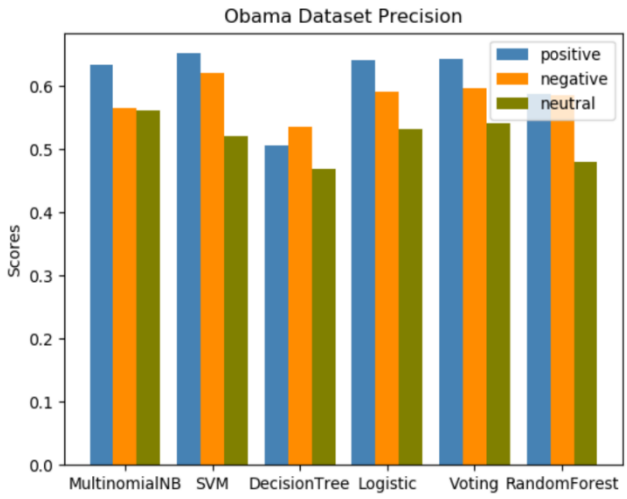
The below figures show the Romney and Obama F1-scores of various classifiers for all the classes. Similar to the Accuracies, Multinominal SVM, Logistic Regression and Voting classifier are the best classifiers which showed significant results than the rest of the classifiers on the Romney and Obama dataset.



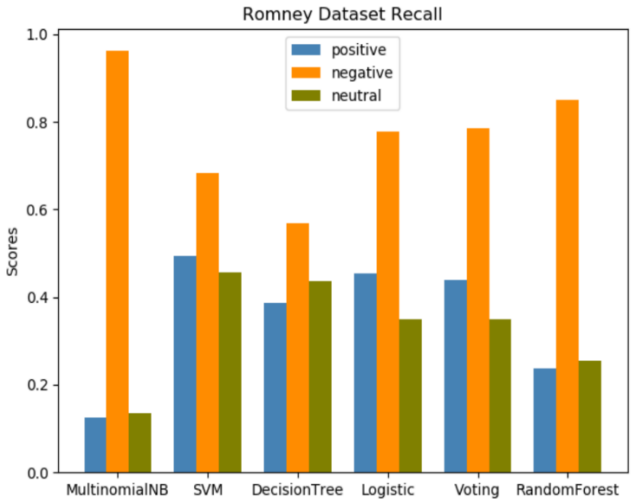


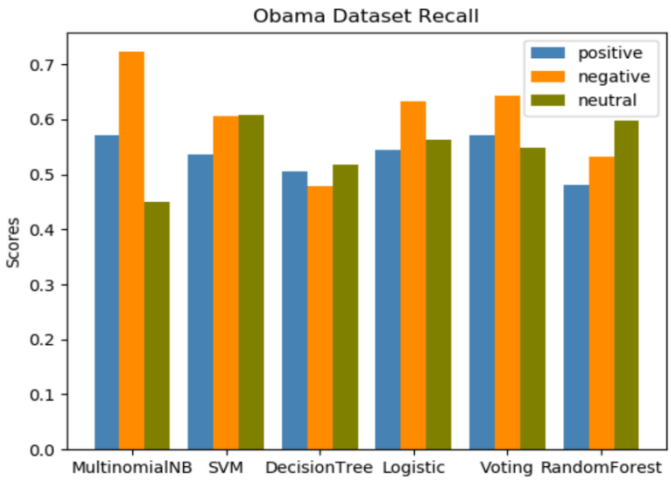
The below figures show the Obama and Romney datasets Precision on various classifiers for all the classes.





Similarly, the below figures show the Obama and Romney datasets Recall on various classifiers for all the classes.





CONCLUSION:

We experimented with several classifiers as shown above. We used k fold cross validation for the training dataset and trained the classifiers. We used SMOTE for the test which balances the skewed data if any and produces better results. We Used unigrams in this project. With many classifiers used in our project, Multinomial Naïve Bayes, SVM and Logistic Regression produced better results than the rest. The Voting classifier is used at the end to see how the ensemble works from the results of the best classifiers. In this project, only the texts of the tweets are considered and other information like the users who tweet them, the times of the retweets and other factors are also potentially useful and as a future scope of this project we would like to experiment with these attributes more and also experiment using n-grams and some other semi supervised classifiers.

REFERENCE:

1. https://scikit-learn.org 2. http://www.nltk.org/genindex.html
3.https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
4. https://imbalanced-learn.org/en/stable/generated/imblearn.over_sampling.SMOTE.html