DATA MINING – SENTIMENT ANALYSIS REPORT

Khushbu Durge(661389838), Swapnil Sagar(677194294)

ABSTRACT:

Sentiment Analysis is the process of 'computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. In this project, we have been provided with a list of tweets from the recent US candidate elections (on Obama and Romney). Twitter here is an online social networking platform where users / account holders can communicate through messages. These messages or "tweets" have been provided to us as the sample data on the basis of which we are required to train our classification model. This sample data consists of tweets which have been classified as "positive", "negative", "neutral" or "neither positive or negative"(1, -1, 0, 2). We first performed data preprocessing on these raw tweets and then ran several classification models to select the best possible fit. We then ran the data on the test set provided later.

INTRODUCTION:

Why is twitter popular for conducting sentimental analysis? According to several surveys conducted, 33% of twitter users post questions of Twitter and many users in turn provide their opinion, review and feedback on the same. This makes Twitter a popular platform for sentimental analysis on recent trending topics. In this project, a set of labelled data has been provided which consists of Twitter feed about the 2012 elections for the US Presidential Elections between the two candidates – Barack Obama and Mitt Romney. These tweets are in CSV format and consist of positive, negative and neutral tweets (respectively for both candidates). This labelled data was used to train classifiers and select the best classification model for both the data sets. We have decided to ignore the class labelled "2" so as to avoid confusion. In this document we cover all the steps performed to build the classification model and then use this model to predict the sentiment (positive, negative or neutral) on the test data.

STEPS: Data Pre-processing, Splitting the data into train, test and validate, build several classification models and calculate the precision, recall, accuracy and F1-Scores for all, Select the best classification models for Obama and Romney tweets respectively.

DATA-PREPROCESSING

As twitter is a social networking site and is not bound by any formal restrictions, there are several usages of internet acronyms, spelling mistakes, emoticons, hyperlinks, unnecessary words or other characters that express special meanings and colloquial slangs by diverse twitter population. Data preprocessing helps in cleaning the raw input data to build a better training model. Considering this project, words or characters like hyperlinks, numbers, question marks, exclamation points, etc. have a less significant contribution to the actual emotion of the tweet. Thus, Preprocessing is the one of the important steps in building an ideal classifier which improves evaluation on parameters.

DATA PREPROCESSING STEPS PERFORMED:

1. Removal of punctuation marks: [!;,?.'"\- %$<>&\(\)\{\}\[\]__=,;:*\\~\+#]
These punctuation marks contribute less to the emotion of the tweet, thus they have been removed.

2. Remove user Reference: Tweets made with the mention of a username (For e.g. : @khushbu) have been edited to remove the username reference. This is because they have no significant contribution regarding the tweet itself.
3. Removal of Hyperlinks: Hyperlinks are like a reference again which is not needed as they just support the actual opinion of the tweet. Hence they have been filtered out.

4. Removal of Stop Words: This has been done while building the TFIDF vectorizer. Words such as "a", "the", etc. have been filtered out as these are stop words and have no real play in predicting the emotion related to a tweet.

5. Removing hashtags and white spaces: Removing hashtags has been a tricky decision. They may or may not contribute to the tweet sentiment. We were able to build a better classifier by removing the hashtags thus this was included in preprocessing. White paces were removed too.

6. Stemming: We used Porter's stemmer (by importing downloading NLTK). Stemming is the process of reducing a word's length to its root word.
What does Stemming really do? Stemming reduces the size of the feature space as many words are originally derived from the same root word. (For example – presidential and president need not be treated as separate words as they mean the same thing).

7. Word Tokenization: We initially decided to perform work tokenization separately, but this was done by using the TFIDF vectorizer itself which we used from the NLTK set.

8. TFIDF Vectorizer: This helped in converting the collection of raw tweets to a matrix of TF-IDF features. This is a statistic that determines the importance of a word in a specific document. In lament terms, it simply assigns a weight to every word and creates a vector that states its importance in a document. This is usually used in text classification tasks.

SPLITTING OF DATA and VECTORIZATION:

Before creating the test and train vectors, we first divided the data into Training and Testing sets.
Once the train and test data were selected and stored in lists, we performed TFIDF vectorization to obtain the train and test vectors and also the list of the total taring data vectors. These vectors were then used to build the below classification models.

CLASSIFICATION MODELS:

Below were the different classification models used and the output obtained (Precision, Recall and F-Scores of positive, negative and neutral data). These are the results obtained by 10-fold cross validation.

1. K-NN CLASSIFIER: We used the K- nearest neighbor classifier initially on the training data. The number of nearest neighbors that we considered was "10". Below were the results obtained on the data sets.

OBAMA

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.44	0.49	0.42
RECALL	0.52	0.30	0.54
F-SCORE	0.47	0.36	0.47

ROMNEY

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.47	0.56	0.33
RECALL	0.33	0.49	0.55
F-SCORE	0.38	0.52	0.41

2**. LINEAR SVM CLASSIFIER**: SVMs are effective in high dimensional spaces. It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient. Here we have simply used the Linear SVM classifier. We can also use the Kernel methods for a data set with more features.

OBAMA

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.56	0.54	0.52
RECALL	0.54	0.55	0.52
F-SCORE	0.56	0.55	0.52

ROMNEY

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.51	0.63	0.42
RECALL	0.40	0.76	0.32
F-SCORE	0.45	0.69	0.36

3. DECISION TREE CLASSIFIER: This is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

OBAMA

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.50	0.54	0.48
RECALL	0.54	0.52	0.47
F-SCORE	0.52	0.53	0.48

ROMNEY

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.39	0.59	0.37
RECALL	0.35	0.63	0.35
F-SCORE	0.37	0.61	0.36

4. MULTINOMIAL NAIIVE BAYES: Naïve bayes classifier for mutiple class label classification.
Naiive bayes predicts the class label of a new text documents by considering prior class probailty **P(c)** and probability distribution of words in the new documents in the learned ducuments.

OBAMA

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.64	0.58	0.55
RECALL	0.58	0.68	0.51
F-SCORE	0.61	0.63	0.53

ROMNEY

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.60	0.59	0.45
RECALL	0.29	0.86	0.24
F-SCORE	0.39	0.70	0.32

5. ADA BOOST: Adaptive boosting algorithm, uses Decision trees as the base classifier, runs base classifier iteratively multiple times giving more weightage to misclassified data every iteration, till there is very low change in the weights. N_estimator used is 200.

OBAMA

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.71	0.55	0.49
RECALL	0.36	0.66	0.60
F-SCORE	0.48	0.60	0.54

ROMNEY

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.50	0.61	0.39
RECALL	0.37	0.69	0.36
F-SCORE	0.42	0.65	0.38

6. LOGISTIC REGRESSION: a regression model where the dependent variable (DV) is categorical. Random state used is 1

OBAMA

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.64	0.59	0.59
RECALL	0.55	0.64	0.55
F-SCORE	0.60	0.62	0.54

ROMNEY

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.59	0.60	0.42
RECALL	0.31	0.84	0.26
F-SCORE	0.41	0.70	0.32

7. XG BOOST: Sklearn implementation of gradient boosting.

OBAMA

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.64	0.57	0.47
RECALL	0.46	0.54	0.64
F-SCORE	0.53	0.55	0.54

ROMNEY

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.60	0.55	0.43
RECALL	0.18	0.91	0.14
F-SCORE	0.28	0.69	0.21

8. RANDOM FORESTS: It is an ensemble method that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes.

OBAMA

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.57	0.57	0.51
RECALL	0.54	0.59	0.51
F-SCORE	0.55	0.58	0.51

ROMNEY

	+1 (Positive)	-1 (Negative)	0 (Neutral)
PRECISION	0.49	0.60	0.41
RECALL	0.30	0.78	0.31
F-SCORE	0.37	0.68	0.35

CONCLUSIONS:

We experimented with several classifiers as shown above. We used 10-fold cross validation to figure out the best possible precision, recall and f1-score results of each classifier. We performed several pre-processing steps to clean the tweets and remove unnecessary stop words, numbers and punctuation marks which helped us improving the overall accuracy of every classifier. We also used n-grams like unigrams and bigrams but decided that "unigrams" gave the best possible results. In this project, text tweets have been considered and information such as referenced users, tweet timings and emoticons have currently been ignored (Although they too may have some contribution in tweet classification). As a future scope of this project we would be analyzing these aspects and exploring more classification models to build a better and more accurate classifier.

BEST CLASSIFIERS:

The best possible results were obtained by using multinomial naïve bayes, support vector machines and the ensemble method – Logistic Regression.
For the test data that we received, we used the Multinomial Naïve Bayes Classifier on the Obama tweets and the Support Vector Machine classifier on the Romney data set.

REFERENCES:

1. Stack Overflow: https://stackoverflow.com/
2. Scikit-learn: http://scikit-learn.org/stable/
3. NLTK: http://www.nltk.org/
4. TFIDF Vectorizer:
http://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html