Research Project Report
CS-583 Data Mining and Text Mining, Fall 2020

Classification of sentiments or opinions expressed in tweets

Nandana Shimoga Prasad & Apurva Raghunath

1. Abstract and Introduction

Twitter is a social networking and microblogging service on which users can post and interact with messages known as "tweets". People who have account in twitter can post, like and retweet tweets. The main purpose it serves is that it connects people and allow people to share their opinions with others.

In this project we are analyzing the sentiments of tweets on twitter relating to 2012 US Presidential elections between Mitt Romney and Barak Obama. These tweets are downloaded from twitter. The sentiments in it can be classified and labelled into 4 classes: positive (1), negative (-1), neutral (0) and mixed (2). In training the data, the mixed class will not be considered. The data will be preprocessed in order to be used for training the model. Data preprocessing is a crucial step here. This is because tweets might contain emojis, acronyms and other special characters which needs to be removed before training and building classifiers. Numerous types of classifiers are used in this project to create and train the model which will classify the new data set of tweets and make the predictions. The labelled data is used to train the classifiers to predict the other tweets and categorize them into positive, negative or neutral label. The positive opinion will be assigned class 1, negative opinion will be assigned class -1 and neutral opinion will be assigned class 0. Each models' performance will be evaluated using metrics such as precision, accuracy, recall and F-1 score. Data preprocessing helps in improving performance metrics of classifiers.

1.1 Sampling of the Training of Data

Data Sampling is a technique used to select, manipulate and analyze subset of actual data in order to identify trends and patterns in the data set. Identifying and analyzing is very cost-effective and more efficient. Sampling helps make data balanced in case the dataset is skewed.

2. Data Pre-processing

a) Removal of punctuations and special characters

The special characters and punctuations like [!@#$%^&*()_+=\?”/`~:’;] are removed and replaced with blank space from tweets as they are not helpful in building the model.

b) Removal of URLs
   The URLs in the tweets are removed and replaced with blank space as they don't help in determining the sentiment of tweet.
c) Removal of HTML tags
   The HTML tags <> are removed and replaced with blank space as they don't help in determining the sentiment of tweet.
d) Removal of usernames
   The usernames are parsed and removed from tweets.
e) Tokenization
   It is the process of breaking text phrase into words called tokens. The list of tokens is the output of data preprocessing and will be used further for building and training the model.

3. Building model with different Classifiers

3.1 Support Vector Machine

Support-Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It constructs hyper-plane or set of hyper planes in a high-dimensional space. The data points which are on the hyper plane are called support vectors. SVM can learn both linear and non-linear decision boundaries. It can be applied to both numerical and categorical attributes. This model gave best evaluation metric results and thus was chosen as the best classifier for testing.

3.2 Decision Tree Classifier

It is one of the predictive model approaches used in data mining. Tree models where the target variable can take a discrete set of values are called classification trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees are very easy to interpret and understand. They are robust to irrelevant and redundant attributes. This model gave least evaluation metric results when compared to others and thus was not chosen.

3.3 Random Forest Classifier

It is a method for 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 or mean/average prediction of the individual trees. It is an ensemble algorithm that considers the results of more than one algorithm of the same or different kind of classification.

3.4 Naïve Bayes Classifier

It is a probabilistic classifier which is built on Baye's theorem. Naïve Bayes assumes that all attributes are conditionally independent. That is, if the class label y is known, then we consider the attributes to be

independent of each other. Higher accuracy levels are achieved as kernel density estimation is employed along with Bayesian network models. This classifier is robust to noise and irrelevant attributes. It is also applicable to categorical and numeric attributes. This model gave second best evaluation metric results.

4. Evaluation

The performance of the classifier models is evaluated using precision, accuracy, recall and F1-score. The evaluation metric results obtained by training the above-mentioned models are tabulated and shown below:

Support Vector Machine					
Obama Positive Class Metrics			**Obama Negative Class Metrics**		
Precision	Recall	F1-Score	Precision	Recall	F1-Score
0.55	0.51	0.53	0.50	0.58	0.54
Romney Positive Class Metrics			**Romney Negative Class Metrics**		
Precision	Recall	F1-Score	Precision	Recall	F1-Score
0.69	0.39	0.50	0.72	0.83	0.77

Overall Accuracy Obama: 0.54 Overall Accuracy Romney: 0.65

Decision Tree					
Obama Positive Class Metrics			**Obama Negative Class Metrics**		
Precision	Recall	F1-Score	Precision	Recall	F1-Score
0.43	0.40	0.42	0.44	0.46	0.45
Romney Positive Class Metrics			**Romney Negative Class Metrics**		
Precision	Recall	F1-Score	Precision	Recall	F1-Score
0.38	0.24	0.29	0.62	0.76	0.68

Overall Accuracy Obama: 0.45 Overall Accuracy Romney: 0.55

Random Forest					
Obama Positive Class Metrics			**Obama Negative Class Metrics**		
Precision	Recall	F1-Score	Precision	Recall	F1-Score
0.47	0.48	0.48	0.56	0.60	0.58
Romney Positive Class Metrics			**Romney Negative Class Metrics**		
Precision	Recall	F1-Score	Precision	Recall	F1-Score
0.76	0.26	0.38	0.68	0.89	0.77

Overall Accuracy Obama: 0.53 Overall Accuracy Romney: 0.65

Naïve Bayes					
Obama Positive Class Metrics			**Obama Negative Class Metrics**		
Precision	Recall	F1-Score	Precision	Recall	F1-Score
0.42	0.62	0.50	0.56	0.53	0.55
Romney Positive Class Metrics			**Romney Negative Class Metrics**		
Precision	Recall	F1-Score	Precision	Recall	F1-Score
0.63	0.32	0.42	0.72	0.79	0.75

Overall Accuracy Obama: 0.51 Overall Accuracy Romney: 0.64

5. Conclusion

We trained the model with several classifiers by tuning various parameters and calculated the performance metrics of each classifier as shown above. In the preprocessing techniques we adapted few steps such as removing the HTML tags, punctuations, usernames, hashtags and URLs which were trivial in building and training the model. After evaluating the performance of each classifiers, we can conclude that SVM classifier model produced the best results with highest F1-score amongst others, followed by Naïve Bayes, Random Forest and Decision Tree classifier models. So, the SVM classifier model was chosen to predict the class labels on the test dataset.

6. References

- http://www.nltk.org/py-modindex.html - NLTK Modules
- https://scikitlearn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html - TFIDF Vectorization
- https://scikit-learn.org/stable/modules - Classifier Evaluation Metrics