Twitter Sentiment Classification

Sachin G Biradar
University of Illinois at Chicago
sbirad@uic.edu

ABSTRACT

Twitter is an online news and social networking service where users post and interact with messages, called "tweets." These messages are restricted to 140 characters. In this project, sentiment analysis is carried on tweets related to the United States presidential election of 2012 between Barack Obama and Mitt Romney.
Labelled data classifying sentiment of tweets as positive, negative, neutral and mixed class are provided for both the candidates separately. A learning model was created using this labelled training data to classify sentiment of any given tweet as positive, negative or neutral class. Various classifiers are used to create the model to classify tweets, their relative performance are discussed in detail. The performance of the model is evaluated by $F_1 score$ and $Accuracy$ of the positive and negative class

1. INTRODUCTION

As a micro blogging and social networking website, Twitter has become very popular and has grown rapidly with several millions of tweets per day posted by millions of users. Thus, sentiment analysis on Twitter is one of the most effective and accurate indicator of public sentiment. In my project, the collected twitter data is related to the United States presidential election of 2012. This data is labelled as positive, negative, neutral and mixed class based on the sentiment of the user. The positive sentiment is labelled as 1, negative as -1, neutral as 0 and mixed as 2. The labelled data is used to train the classifier to predict the sentiment of new tweets.
As twitter is a social networking site the language is not bound by any formal restrictions, there tends to be usage of internet acronyms, spelling mistakes, emoticons, other characters that express special meanings and colloquial slangs by diverse twitter population. Hence preprocessing is to be done on the data set before training and building a classifier. Preprocessing is the one of the important steps in building an ideal classifier which improves the evaluation parameters. It is not necessary that equal number of people have positive and negative opinion hence the given labelled dataset can skewed. In the training dataset provided in the project, Obama dataset is fairly balanced as shown in the pie charts with similar proportions of the positive, negative and neutral datasets whereas Romney dataset is imbalanced with negative dataset dominating the other two datasets. Hence the given data has to be oversampled to adjust the class distribution of a data set. The cleaned and sampled dataset is then converted to feature vectors which are used to train the model. Several classifiers were tried and evaluated using 10-fold cross-validation

2. TECHNIQUE

I will discuss the techniques in the following sections. Before we dive into the details of the algorithms, let us take a look at the Twitter data first and discuss its characteristics.

2.1 Twitter Data

Twitter has developed its own language conven-

tions. The following are examples of Twitter conventions.[1]

- *"RT"* is an acronym for retweet, which is put in front of a tweet to indicate that the user is repeating or reposting.
- *"#"* called the hashtag is used to mark, organize or filter tweets according to topics or categories.
- *"@username"* represents that a message is a reply to a user whose user name is username.
- *Emoticons* and *colloquial expressions* are frequently used in tweets, e.g. ":-)", "lovvve", "lmao".
- *External Web links* (e.g. http://amze.ly/8K4) are also commonly found in tweets to refer to some external sources.
- *Length*: Tweets are limited to 140 characters. This is different from usual opinionated corpora such as reviews and blogs, which are usually long.

2.2 Preprocessing

I removed retweet string RT as it does not add value to the sentiment of the tweet. External links and user names (signified by @ sign) are eliminated. I restored popular contractions to their corresponding original forms using a dictionary of contractions (e.g. "cant" to "can not"). A predefined set of hashtags which are used to mark topic as trending were removed. Hashtags were split into separate meaningful words. Incorrect spellings were corrected. All the words were then replaced by their stem words.[2]

2.3 Oversampling

The given has approximately balanced class distribution for Obama's tweets whereas for Romney's tweet the class distribution is skewed. Hence the classifier maybe biased towards one class if trained without sampling hence I sampled the data using *SMOTE*: Synthetic Minority Oversampling Technique.[3] Given below is the class distribution before and after oversampling. all the classes have equal probability distribution are same for all the classes after oversampling

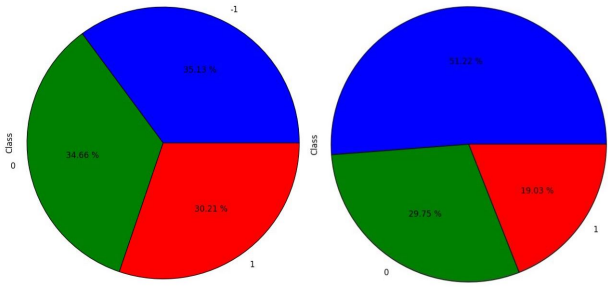

Figure 1: Class distribution for Obama and Romney before sampling

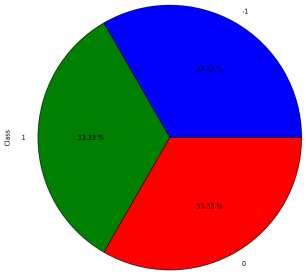

Figure 2: Class distribution for Obama and Romney after sampling

2.4 Feature Generation

After the data is cleaned and sampled it was then converted to vector format by using *if-idf* (term frequency-inverse document frequency) vectorizer with minimum document frequency of 0.00125, maximum document frequency of 0.7 and n-gram range of (1, 5)

2.5 Model training

The generated features were fed to the following classifiers-[4]

- **Multinomial Naive Bayes:** MultinomialNB implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate multinomial distributions; ie with a multinomial event model, samples (feature vectors) represent the frequencies with which certain events have been generated by a

multinomial ($p_1,\ldots,p_n$) ($p_1, \ldots, p_n$) where p_i p_i is the probability that event i occurs (or K such multinomials in the multiclass case). A feature vector x= ($x_1,\ldots,x_n$) $x=(x_1,\ldots,x_n)$ is then a histogram, with x_i x_i counting the number of times event i was observed in a particular instance. The results of this classifier is one of the best among many classifiers employed in the project.

- **Random Forest:** 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 havent been consistent.
- **Logistic Regression:** Logistic Regression model is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. The idea behind this model is that one should prefer the most uniform models that satisfy a given constraint. After MultinomialNB, logistic Regression is the classifier which produced accurate and consistent results.
- **Decision Tree:** 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). Life Random Forest the results of decision tree havent been consistent.
- **Stocastic Gradient Descent:** This estimator implements regularized linear models with stochastic gradient descent (SGD) learning: the gradient of the loss is estimated at each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate). SGD allows minibatch (online/out-of-core) learning, see the partial fit method. For best results using the default learning rate schedule, the data should have zero mean and unit variance.
- **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 determines the best classifiers and averages the evaluation parameters obtained from those best classifiers.

3. EVALUATION

The evaluation parameters of the classifiers are F-Score of positive and negative class and the overall accuracy. The below chart shows evaluation metrics for various classifiers for both Obama's and Romney's tweet.
For both Obama and Romney, Multinomial NB and Logistic Regression are the best classifiers which showed significant results than the rest of the classifiers

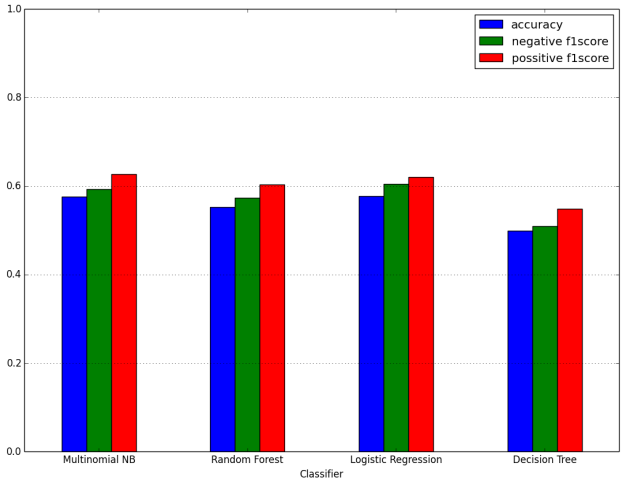

Figure 3: Evaluation for Obama's tweets

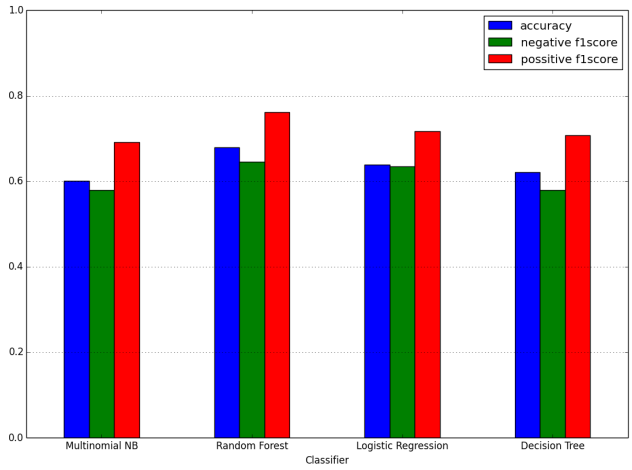

Figure 4: Evaluation for Romney's tweets

4. CONCLUSION

I experimented with several classifiers as shown above. I used 10-fold cross validation for the training dataset and trained the classifiers. I used SMOTE(Synthetic Minority OverSampling technique) for the test which balances the skewed data if any and produces better results. The results with and without SMOTE is quite drastic. I also tried many steps in the preprocessing which might improve the results such as lemmatization and splitting hastags into distinct words. These changes had minor impact on the final results. The ngrams such as unigram and bigram were tried but they are decreasing the performance and hence were not used. With many classifiers used in my project, Multinomial Nave bayes and Logistic Regression produced better results than the rest. The Voting classifier is used at the end to produce the average of the results from 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 I would like to experiment with these attributes and few more supervised learning optimization algorithms such as Neural networks and some semi supervised learning models.

APPENDIX

A. REFERENCES

[1] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. 2011.

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.

[3] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.