Analyzing Different Techniques for Sentiment Analysis on Twitter Data (Demonetization in India)

Mansha Imtiyaz
mansha.imtiyaz@mail.mcgill.ca

Abstract

Sentiment Analysis is the computational treatment of opinions, sentiments and subjectivity of text. This report will review and evaluate some of the various techniques used for sentiment analysis. Different classification algorithms such as Naive Bayes, Logistic Regression and an artificial neural network technique - Doc2Vec has been implemented and compared on the twitter data set for best performance. Logistic Regression using doc2vec model performs comparatively better. Also, the above techniques and NLTK's Vader will be used to analyze the sentiments of tweets related to aftermath of currency demonetization in India.

1 Introduction

Sentiment Analysis (SA) is the process of determining whether a piece of writing is positive, negative or neutral. It is extremely useful in social media monitoring as it allows us to gain an overview of the wider public opinion behind certain topics. Many companies, political parties and different brands have used SA to gauge public reaction.

Sentiment Analysis is a widely discussed topic for research and much of the work has been done in this field. To automate sentiment analysis, different approaches have been applied to predict the sentiments of texts. Medhat et al. (2014) in their paper explore the sophisticated categorizations of a large number of recent articles and survey the recent trends of research in the sentiment analysis field. Ruby and Mike (2009) combine rule-based classification with classification algorithms and show that hybrid approach can improve the classification effectiveness. Hutto and Gilbert (2015) proposed an algorithm (VADER) and suggested a method of analyzing social media text using a rule-based model. They even compared its effectiveness with eleven state-of-practice benchmark tools and found that it performs favorably across contexts compared to any other benchmarks.

This report draws upon the existing approaches used for sentiment analysis and compares the effectiveness of each approach. This will look at different models for classifying "tweets" into positive and negative sentiment by building models for classification tasks. For this, NaiveBayes and Logistic Regression models are used to train the data. Two types of feature models are used: bag of words model with weight distribution of positive to negative tweets and a deep learning technique for classifying features (doc2vec). Once the models are trained and tested, sentiment analysis of tweets obtained by live streaming of data from twitter using hashtags (2011) is done and the results are compared for effectiveness among each other and with the rule-based model (VADER).

In the last section, analysis is being shown using the model that performs the best in the above cases. I have focused on the topic of currency demonetization. The demonetization of 500 and 1000 banknotes was a step taken by the Government of India on 8 November 2016, ceasing the usage of all 500 and 1000 banknotes. After the announcement was made, there was a lot of chaos and mixed reactions from everyone. This affected the country as a whole. In this report, the public reaction towards

this change on social media will be analyzed.

2 Data Collection

For training the classification models, Stanford Twitter Data-Set (Sentiment140,) was used. The data-set has been divided into four files - positive and negative tweets for training and testing the model respectively.

For analyzing the sentiments towards demonetization, live twitter feed was streamed using Twitter API filtering the search to track the following keywords- 'Demonetization','#IndiaFightsCorruption', '#Modidemonetizationcircus' and '#IndiaDefeatsBlackMoney'.

2.1 Data Processing

Pre-processing the data is the process of cleaning and preparing the text for classification. The data collected contains lots of noise and uninformative parts such as HTML tags and scripts. In addition, on word level, many words in the text do not have an impact on the general orientation of it or the polarity, such as stop words, punctuation, etc. Due to the particularity of data towards Indian audience, the data contained many non-english words, some of which were removed after manual cleaning. This would have some effect on the final accuracy, as the model wasn't trained on a domain specific data set.

Following steps were taken for processing data: online text cleaning, white space removal, stemming, removing emoticons, stop words removal and removing non-ascii words. From the 8000 tweets collected for analysis of sentiments regarding demonetization, 5876 were extracted after processing and cleaning.

3 Models

Two classification models - Naive Bayes and Logistic Regression are trained on the dataset using two feature vector models described below. The results are stored and compared. Third Model - NLTK's Vader is also used to calculate the polarity of the tweets and the results are analyzed. The two feature vector models used are explained below.

3.1 Model 1- NLP Feature Vectors

For classification experiments, variety of features have been used. For the baseline, we use bag of words (unigram) model. I have also included features typically used in sentiment analysis, namely features representing information from a sentiment lexicon. Also, features to check if the word is in at least 1 percent of the positive texts or 1 percent of the negative texts and if it is in at least twice as many positive texts as negative texts have been used.

3.2 Model 2- Doc2Vec Model

Second feature vector model used is actually a deep learning technique that is a build up on Word2Vec. Word2Vec captures the context of words with the aim to predict a given word using the surrounding words or to predict a window of words given a single word. However, it ignores the word order.

Mikolov and Le (2014) proposed paragraph vector feature, that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. These feature vectors had been used to train movie dataset and the accuracy was near to 86%.

Distributed Memory(DM) model of Doc2Vec has been used in the current project to train the dataset. DM attempts to predict a word given its previous words and a paragraph vector. Since its accuracy is exceptionally good for the IMDB Movie Review data set, the expectation is that it will perform better than any of the other methods on the twitter dataset as well.

Gensim's Doc2Vec implementation has been used. It requires each tweet to have a label associated with it. This has been done using the LabeledSentence method. After instantiating the DM models, vocabulary is build over all the sets (train, test and also the streamed twitter dataset). On running the code, we pass through the data set multiple times, shuffling the training each time to improve accuracy.

3.3 Rule Based Model

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis library. For the purpose of the project, SentimentIntensityAnalyzer function was used to assign

compound polarity to each tweet. This method has been used for final analysis of the twitter data regarding demonetization and some results have been explained (see section 5).

4 Results and Discussion

The first step was to train the models using the training data. Table 1 shows the results which were obtained after training and testing on Sentiment140 data.

Model	Naive Bayes	Logistic Regression
Accuracy	53.27%	60.3%

Table 1: Accuracy with NLP Feature Vector Model

Using the doc2vec feature model, the results were as follows:

Model	Naive Bayes	Logistic Regression
Accuracy	55.47%	65.7%

Table 2: Accuracy with doc2vec Model

Logistic Regression using doc2vec model performs comparatively better than the other models. However, the results weren't as good as expected and not much improvement over the baseline method. The reason could be that the informal nature of tweets affected the accuracy. Also, since doc2vec has been proven to give good results on movie review database, length of the text could also be a factor as more words provide more features to be trained on, and hence, better accuracy.

Model	VADER
Accuracy	77.79%

Table 3: Accuracy with VADER

For comparison, the test data was also analyzed using NLTK's Vader library, taking polarity from -0.1 to 1 as *positive* and -1 to -0.2 as *negative*. It gave a good accuracy of 77.79% which is a lot better than the above two models.

For the second step, tweets regarding demonetization were analyzed using the LR Model of both feature vector models.

Table 4 clearly shows that for the given dataset, doc2vec model performed better as majority of the tweets have been classified as positive for Model 1. On checking the results as well, we come to know

Model	Positive Values	Negative Values
Model 1	5552	324
Model 2	2234	3642

Table 4: Predicting sentiments for Demonitization twitter data

that the tweets which are clearly negative have been labeled as positive. For demonstration, lets consider two tweets:

Tweet 1 : *Sorry to mention I'm not going with ur option I will go with "None of the above" or "Modi a foolish PM of India"*

Tweet 2 : *Demonetisation has "failed" to unearth a single penny of black money in the last one month.*

Model	Tweet 1	Tweet 2
Model 1	Positive	Positive
Model 2	Negative	Negative

Table 5: Sentiment Analysis of given tweets

Normally reading over the tweets, we would know that both tweets have negative sentiments which the doc2vec model has correctly guessed.

However , the second model doesnt always give accurate results as can be seen from the following tweet.

I salute the people of India for wholeheartedly participating in this ongoing Yagna against corruption, terrorism

The model labels the above tweet as positive. This may be due to conflicting terms - 'salute' which is positive and 'corruption', 'terrorism', which are negative. But as a whole, doc2vec model performs better than the first model, with VADER outperforming both.

5 Demonetization in India- Brief Analysis

One of the limitations of using classification algorithm used above is that the data set that we have used is binary classified. Had the dataset included neutral label as well, the results could have been more interesting. So, for analysis of tweets regarding demonetization, we will use VADER. NLTK's Vader gives polarity in a range of -1 to +1, which can then be classified accordingly in three or more classes. Three classes were chosen, with -1 to 0 considered as negative, 0 as neutral and 0 to +1 as positive.

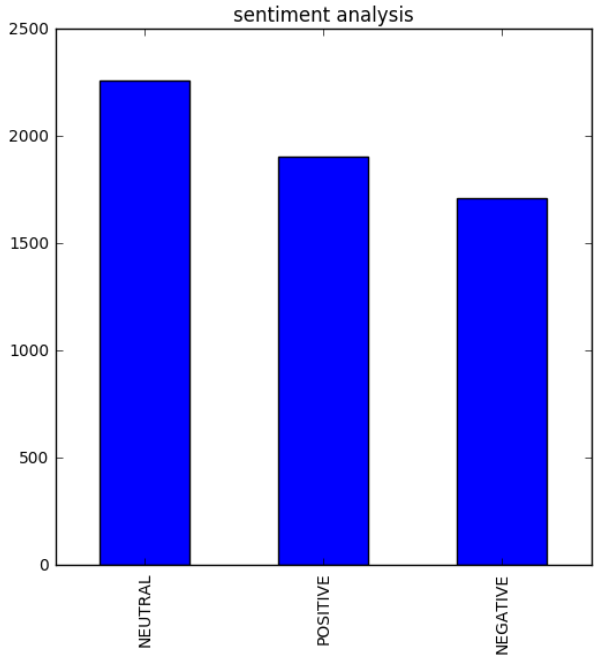

Figure 1: Distribution of Polarity of Tweets

The polarity distribution of tweets has been show in Figure 1. By testing on all of the models, it is clear that positive sentiment is more than the negative sentiment, however by a small proportion, and majority have stayed neutral to the change.

Hashtag	Count
#modidemonetisationcircus	944
#indiadefeatsblackmoney	705
#blackmoney	503
#indiafightscorruption	294

Table 6: Hashtag Distribution

The count for the hashtags used in the tweets is shown in Table 5. This shows that those who were tweeting in support of the demonetization (#indiadefeatsblackmoney) were almost similar in numbers to ones pointing to demonetization as a circus.

On forming the word clouds for most commonly used terms, we see that the words - angry, businessman, shock, poor, protest and economy are at top. Checking the term co-occurrence for words such as 'economy' and 'poor', we get the wordclouds as in Figure 2 and Figure 3 respectively. We can predict from this that people are tweeting that economy has been terribly affected, black and white money, cashless economy, etc. Same analysis can be drawn for the term 'poor' from Figure 3.

Figure 2: WordCloud For Term Co-Occurrence with 'Economy'

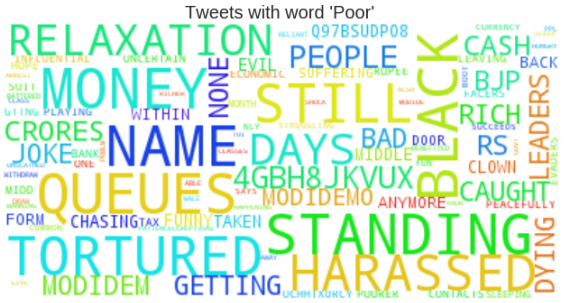

Figure 3: WordCloud For Term Co-Occurrence with 'Poor'

6 Conclusion

From the above results, we can conclude that logistic regression using doc2vec performs better than the nlp feature vector model with naive-bayes and/or logistic regression, as had been expected. Though, as seen in some cases, it isn't always accurate. But the dataset uses only positive and negative labels. In real world applications, neutral tweets cannot be ignored. Proper attention needs to be paid to neutral sentiment. Better data-set to train the model could be used with labels with domain-specific tweets which would have accounted for the non-english terms as well. On the other hand, VADER gave a good accuracy of 77% for the same dataset compared to the other models.

Also, analysis of the tweets regarding demonetization shows that there are no extreme sides which are against or who support the change. Though, currency demonetization has created a lot of chaos, majority of people are taking it well and looking at it as

a big step towards positive change.

References

Eric Gilbert C.J. Hutto. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence.

Johanna Moore Efthymios Kouloumpis, Theresa Wilson. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence.

Tomas Mikolov Quoc V. Le. 2014. Distributed representations of sentences and documents. *Proceedings of The 31st International Conference on Machine Learning*, pages 1188–1196.

Mike Thelwall Rudy Prabowo. 2009. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157.

Sentiment140. Stanford twitter set. *http://help.sentiment140.com/for-students*.

Hoda Korashy Walaa Medhata, Ahmed Hassan. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.