

Sentiment Analysis of Twitter Data on Demonetization Using Machine Learning Techniques

N. M. Dhanya^{1(✉)} and U. C. Harish²

¹ Department of Computer Science and Engineering, Amrita School
of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
nm_dhanya@cb.amrita.edu

² Guruvayurappan Institute of Management, Coimbatore, India
harishuc@gmail.com

Abstract. Social media like twitter and Facebook is seen as a space where public opinions are formed in today's world. The data from these tweets and posts can provide valuable insights for policy makers and other agencies to propose and implement policies better. An attempt is made in this paper to understand the public opinion on the recently implemented demonetization policy in India. A sentiment analysis is carried out on twitter data set using machine learning approaches. Twitter data from November 9th to December 3rd is considered for analysis. The data set is pre-processed for cleaning the data and making it possible for analysis. A final set of 5000 tweets are analysed using machine learning techniques like SVM, Naïve Bayes classifier and Decision tree and the results are compared.

Keywords: Twitter data · Sentiment analysis · Machine learning
Demonetization

1 Introduction

Micro blogging sites such as twitter and Facebook are very powerful communication tools in the current highly interconnected world. Millions of messages are appearing daily in these media. The users can write about their life events, share their opinion and discussion on current issues can be done through this media. In the current scenario where the internet speed is very promising, these media act as an easy communication tool for spreading ideas and developing opinions. Any information such as new product launches, opinion polls [1] and current issues can be posted in these media to get responses of the public. The data collated from social media can be successfully utilized for analyzing people's opinion and sentiments. This analysis can be effectively used for product marketing.

1.1 Applications of Sentiment Analysis

Sentiment analysis has so many applications in real life across various domains.

E-Commerce: The most popular analysis that can be done using social media sentiment analysis is the e-commerce application. An e-commerce website allows users to submit reviews of the product which they have purchased. These reviews can be based on product quality, mode of delivery and overall customer satisfaction. Other customers can see the reviews and based on the review they can decide on the purchase. For the company management these reviews can be analyzed to make their products better. Based on the negative comments the products can be improved.

Reputation analysis: Reputation analysis can be done by using micro blogging sentiment analysis. Similar to customer analysis, reputation management is also equally important for management. The analysis helps them to manage and strengthen the brand reputation. Sentiment analysis helps companies to analyze how their products are being perceived by the online community.

Election exit polls: Sentiment analysis can also be used by government to analyze the strength and weakness of various government policies by analyzing the opinions of the public. The sentences like “The minister himself is corrupted, then how do you expect the truth will come out” clearly shows the negative sentiment and “What a nice policy, which will enhance the country’s growth” specifies a positive sentiment towards the government. The polls like the one shown in Fig. 1 can be used by candidates to predict the votes. The opinion and the place from where the opinion is posted can be further analysed and used in campaigning.



Fig. 1 Sample Twitter poll

In India, a very significant policy initiative of demonetizing Rs. 500 and Rs. 1000 notes was announced on November 8, 2016 by the Prime Minister, Shri Narendra Modi. In this paper, the sentiment analysis of this policy is carried out using the most popular micro blogging site, twitter. In twitter, users post small lines of comments called tweets to mark their opinion. The following are the characteristic features of tweets.

1.2 Tweet Features

Length of a Tweet: The maximum length of text that is possible in twitter is 140 characters. Even though it cannot be a meaningful grammatically correct sentence, it can be considered as a single entity or a sentence. Normally in product reviews the sentences will be more structured and analysis will be much easier than a tweet.

Language: Tweets can be given through a variety of languages. Tweets will mostly be in colloquial languages filled with slangs and misspellings.

Hashtags: A Twitter hashtag is simply a keyword phrase, spelled out without spaces, with a hash sign (#) in front of it. For example, #inboundchat and #ILoveChocolate are both hashtags. The use of hashtag is gaining popularity in tweets. Analysis shows that almost all tweets contain hashtags. Hashtag campaigns are conducted for increasing the popularity of a product. Figure 2 shows an example.

Data availability: Another feature of tweets is the availability. With the Twitter API, in several languages it is possible to collect millions of tweets for training. Here we used python API for data collection. There are also some datasets [2, 3] available that have automatically and manually labelled tweets.



Fig. 2 Hashtags in Twitter

2 Related Work

Sentiment analysis can be considered as a natural language processing task, classifying the sentences into different clusters called positive, negative and neutral. It has many granularities right from the document level to the sentence level [4]. Micro blogging sites such as twitter and Facebook is updated with real time reactions and opinions. There has been large amount of prior research in sentiment analysis mostly in the domain of product reviews, movie reviews and election exit polls.

A distant supervision technique is introduced by Pak and Paroubek (2010) which automatically collects the dataset from the web [5]. They used happy and sad emoticons for positive and negative sentiments. The tweets from “New York Times” and “Washington Post” and other popular news paper accounts are collected. These tweets are used as training sets of subjective tweets. The techniques such as Unigrams and filtered n-grams are used for their sentiment classification. The negative sentiments are considered by attaching negation words like “no”, “not” to the words before and after them. These negation words along with positive word can be considered as negative sentiments.

The sentiment analysis of tweets is considered to be different task as the tweets consists of informal language which are also sometimes new and creative [6]. Koulompis, Wilson and Moore (2011) use the earlier work done in hashtags and sentiment analysis to build their classifier. The most frequent hashtags are identified using Edinburgh Twitter corpus and they manually classify these hashtags. These are in turn used to classify tweets. They have used n-grams and part-of speech features. In addition, they have built a feature set from Internet lingo dictionary and MPQA subjectivity lexicon. Their analysis identifies that usage of Part-of-Speech causes a reduction in accuracy whereas n-gram features with lexicon features gives better results.

Another approach is discussed by Saif et al. [7]. They introduce a semantic based approach which identifies the items in a tweet, like a person, organization etc. [7]. The effect of stop words in the sentence is studied and they concluded that removal of such stop words may cause undesirable effect in the meaning of the sentence.

Identifying negation words is a technique usually studied in sentiment analysis. The words like “not”, “no” and “never” can change the meaning of the sentence to the opposite direction. That is positive sentences can be changed to negative and negatives can be changed to positives. Because of such words the nearby word meanings may become opposite. For example “not good” change the meaning of a positive word “good” to negative. Thus negation words play an important role in sentiment analysis, which may change the entire meaning of a sentence.

The scope of negation words can be considered to be from the point of appearance to the next punctuation. This scope is analyzed in [8]. The scope is analyzed for the sentiment of that sentence. They identified explicit negation cues for each word in the sentence. The distance between the word and the nearest negative word from left and right is calculated and the sentiment score is calculated.

The adverse effect of drug [9] is analyzed in this work based on the drug related tweets. A pipelined algorithm is proposed which uses a simple drug-related classification and sentiment analysis to extract Adverse Drug Events on Twitter. The SVM and Naïve bayes classification is used for comparing the results and SVM is getting higher accuracy.

Bao et al. [10] compared the performance of different pre-processing approaches for twitter data. They compared the impact of pre processing on the accuracy of sentiment analysis. The effects of URLs on the tweets, negation words, repeated letters, stemming [11] and lemmatization were evaluated and conclude that removal of all these improves the accuracy significantly.

All the above stated techniques rely on n-gram [12] features. To improve the accuracy significantly some authors used feature selection also. In this paper we are creating a sample dataset from the tweets. The training data is created with the analysis on the tweets. By doing pre processing and feature extraction the important features are extracted and using machine learning approaches a classifier model is created.

3 Demonetization

In a dramatic announcement on the evening of 8th November, 2016, Mr. Narendra Modi, the Prime minister of India announced that currency notes of Rs. 500 and Rs. 1000 will no longer be a valid tender from 12.00 midnight. He announced that citizens could exchange or deposit the invalid notes at the banks till 30th December 2016. Also a number of restrictions on withdrawal of money from ATMs and banks were implemented. In his announcement he made it clear that this is an unavoidable step to eradicate fake notes and to curb black money which was in circulation in the country.

During the days that followed huge queues were seen in front of banks and ATMs and media reported that huge difficulties were faced by citizens due to the restrictions imposed. The opposition parties raised the issue in the parliament and parliament functions got disrupted many times.

There were also reports that the common citizens of the nation supported the move as they believed that this is an important step towards bringing out black money. The government also projected this move as a significant step towards making India largely a cashless economy where most of the transactions are made in electronic/digital form. The social media is also abuzz with discussions about the benefits and difficulties faced from the decision.

The decision and implementation of demonetization is seen as a historic decision with long term ramification in political and economic future of India. This analysis using tweets is an attempt to understand the general sentiment among the public about the demonetization policy implemented by the Government of India.

4 Dataset

Data set for the analysis is collected from twitter API methods for a period of 25 days. Around 12,500 tweets were collected over a period from 09-11-2016 to 03-12-2016. The data under the hashtag of demonetization and black money are collected and a live streaming of data on the day of 03rd December is also considered.

This data is pre processed and unwanted tweets are removed and a collection of 5000 tweets with two different topics, namely, demonitization and black money is collected and classified. The corpus which we collected contains the following information. Each entry contains a tweet id, the tweet text and the sentiment label. Python libraries for twitter can be used to collect the information about the tweets like creation date, creator name, etc. Each tweet is classified into three categories such as positive, negative and neutral. The irrelevant tweets are also classified as neutral. Positive is used for showing positive sentiment or positive opinion towards demonetization, Negative for showing negative sentiment towards the movement and Neutral for non English text or off-topic comments. The following table shows sample tweets with label (Table 1).

Table 1 Tweets in corpus

Label	Example
Positive	Very good way to handle current situation strategy demonitization
Negative	This demonitization has become such a mess that even Bhakts in my family have stopped defending it
Neutral	Narendramodi will respond in both the houses of parliament

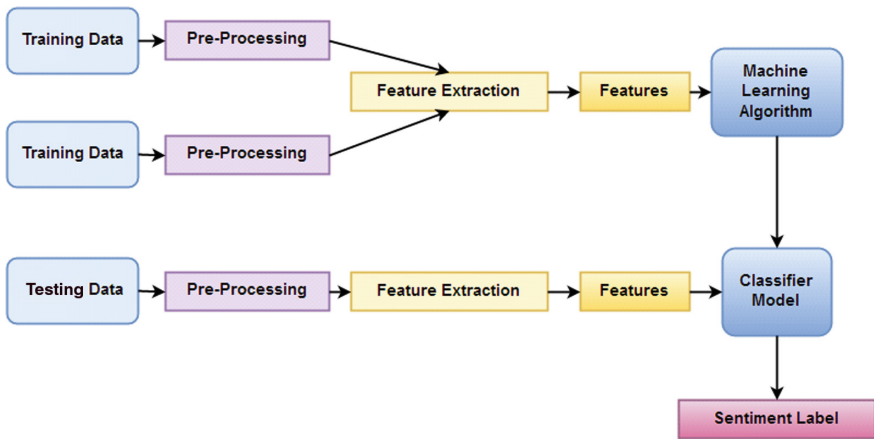


Fig. 3 Sentiment analyzer architecture

5 Methodology

A sample dataset with the labels of positive, negative and neutral is created as training dataset. The data is pre-processed to eliminate some of the unwanted features from the tweets. Different machine learning classifiers are used to find the best sentiment analysis (Fig. 3). Various pre-processing steps such as removing punctuations, emoticons, twitter specific terms and stemming are done and the data is cleaned for analysis. The machine learning techniques such as Naïve Bayers classifier, Decision tree and SVM are used for classification and the results are compared.

Pre Processing

The data generated from twitter is usually not suitable for learning or analysis directly. Hence the data should be normalized to make it in a better format before applying any techniques. Here we are using pre-processing techniques which will eliminate unwanted text from the tweets and hence reduce the features. Thus the data is made suitable for all the learning algorithms. Stemming is also used for pre-processing the data. The porter stemmer is used as stemming algorithm.

Figure 4 illustrates various features seen in micro-blogging. There can be Re tweet symbol, Special characters such as !,.. Emoticons, punctuations and handles. Our pre



Fig. 4 Sample tweet

processing module removes all these unimportant data from the tweet texts and makes it ready for feature extraction (Tables 2, 3, 4, 5 and 6).

Features

There can be a large variety of features based on which a tweet classification can be done. The most popularly used and basic feature set is word n-grams [12]. However, domain specific information present in tweets can also be used for classifying them. A sample dataset with sentiment classification is created. This is done by preparing a positive word corpus, which can appear as positive or words that are agreeing with the demonetization movement like awesome, good, hats off (awesome movement, good decision, and hats off to modiji). The negative word corpus is also created in a similar way like shocking, mess and slows down (Shocking decision, such a mess and slow down the economy rate). For each tweet a sentiment score is calculated from the positive and negative word corpus (Fig. 5; Table 7).

The sentiment scores are calculated on the basis of words used in the tweets. The calculation is shown in Table 7. Each tweet is evaluated and a numeric score is calculated. Based on this score, sentiment labels are attached according to the following rules.

If the score is less than 0 (negative values) a sentiment of negative is attached to the tweets and if the score is greater than 0 (positive) a positive sentiment is attached and finally if the score is 0 it is classified as neutral.

Table 2 Bernoulli naive Bayes

Tweets	Precision	Recall	F1 score
Negative	0.90	0.24	0.38
Neutral	0.64	0.99	0.78
Positive	0.83	0.26	0.39
Avg/total	0.74	0.67	0.61

The accuracy score is 66.75%

Table 3 Multinomial naive Bayes

Tweets	Precision	Recall	F1 Score
Negative	1.00	0.19	0.31
Neutral	0.61	1.00	0.76
Positive	1.00	0.12	0.22
Avg/total	0.78	0.63	0.54

The accuracy score is 63.25%

Table 4 SVM

Tweets	Precision	Recall	F1 Score
Negative	1.00	0.59	0.74
Neutral	0.79	1.00	0.88
Positive	0.95	0.84	0.83
Avg/total	0.87	0.84	0.83

The accuracy score is 84.00%

Table 5 LDA

Tweets	Precision	Recall	F1 score
Negative	0.32	0.68	0.44
Neutral	0.81	0.70	0.75
Positive	0.89	0.41	0.56
Avg/total	0.74	0.62	0.64

The accuracy score is 62.25%

Table 6 Decision tree

Tweets	Precision	Recall	F1 score
Negative	1.00	0.19	0.31
Neutral	0.61	1.00	0.76
Positive	1.00	0.12	0.22
Avg/total	0.78	0.63	0.54

The accuracy score is 63.25%

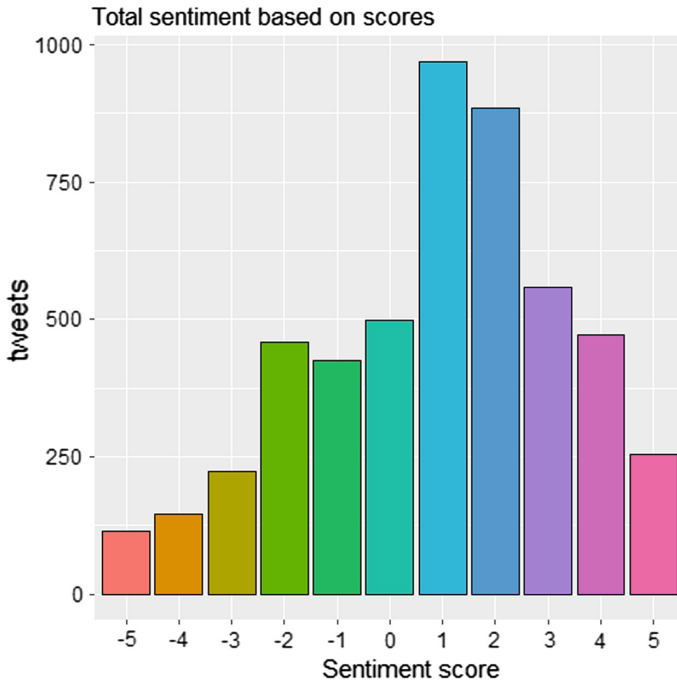


Fig. 5 Sentiment score

Table 7 Sentiment score

Tweets	Score
“Demonetization is brilliant. But people are suffering a bit”	0 (+brilliant, -suffering)
“Kejriwal posts pic of dead robber and claims it’s #demonetization related death? How shameless has this man become?”	-3 (-dead, -death, -shameless)
“Putting Nation over Party Politics nitishkumar supports PM narendramodi on Demonetization”	+1 (+Supports)
“A man shaved his head at Jantar Mantar in protest against Demonetization”	-1 (-against)

6 Results

The result proves that positive sentiment score is more compared to the other sentiment scores, concluding that support for demonetization is high. There are also people who feel negative about it.

The following tables show the accuracy measures for different classification methods which are used in this paper. All the machine learning techniques are written in Python code with the help of NLTK [13] and sklearn packages. The data

preprocessing and labeling is also done with Python [14] packages and a predefined positive and negative corpus.

The reduction in the accuracy rate is due to the Hindi and Tamil tweets which are present in the dataset and majority of these are classified as neutral. This can be improved by adding Hindi and Tamil words to the corpus with some technologies [15, 18] (Fig. 5).

7 Conclusion

Sentiment analysis on twitter dataset is done on the recent issue of demonetization. The analysis shows largely a positive opinion on the move though negative opinion is also present. Various machine learning algorithms [16] for classification are used for the results comparison. For our dataset, SVM shows the maximum accuracy. The major disadvantage of our approach is that many tweets are written in local languages like Hindi. The Hindi words are written in English language and we didn't include any such words in the positive and negative corpus. Because of this all tweets in Hindi, written in English are classified as neutral. In future those words can also be included and accuracy can be increased for analysis. As future work similar techniques like [17] can be used to reduce the neutral count.

References

1. Gokulakrishnan, B., Priyanthan, P., Ragavan, T., Prasath, N., Perera, A.: Opinion mining and sentiment analysis on a Twitter data stream. In: 2012 International Conference on IEEE, Advances in ICT for Emerging Regions (ICTer) (2012)
2. Go, A., Bhayani, R., Huang, L.: Twitter Sentiment Classification Using Distant Supervision. Technical Report (2009)
3. Sanders, N.: Twitter Sentiment Corpus. <http://www.sananalytics.com/lab/twitter-sentiment/>. Sanders Analytics
4. Sun, S., Luo, C., Chen, J.: A review of natural language processing techniques for opinion mining systems. *Inf. Fusion* **36**, 10–25 (2017)
5. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. European Language Resources Association(ELRA), Valletta Malta (2010)
6. Kouloumpis, E., Wilson, T., Moore, J.D.: Twitter sentiment analysis: the good the bad and the Omg! ICWSM 538–541 (2011)
7. Saif, H., He, Y., Alani, H.: Semantic Sentiment Analysis of Twitter. In: The Semantic Web-ISWC 2012, pp. 508–524. Springer, Berlin (2012)
8. Councill, I.G., McDonald, R., Velikovich, L.: What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In: Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pp. 51–59. Association for Computational Linguistics (2010)
9. Peng, Y., Moh, M., Moh, T.S.: Efficient adverse drug event extraction using twitter sentiment analysis. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2016)
10. Bao, Y., Quan, C., Wang, L., Ren, F.: The role of preprocessing in Twitter sentiment analysis. *Intell. Comput. Methodologies* 615–634 (2014)
11. Smirnov, I.: Overview of Stemming Algorithms. Mechanical Translation (2008)

12. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retrieval* **2** (1–2), 1–135 (2008)
13. NLTK (Nature Language Tool Kit): Last Retrieved on 21 March 2015 from <http://www.nltk.org/>
14. Bird, S., Klein, E., Loper, E.: *Natural language processing with Python*. O'Reilly Media, Inc. (2009)
15. Nivedhitha, E., Sanjay, S.P., Anand Kumar, M., Soman, K.P.: Unsupervised word embedding based polarity detection for tamil tweets. *Int. J. Control Theor. Appl.* **9**, 4631–4638 (2016)
16. Quinlan, J.R.: *C4. 5: Programs for Machine Learning*, vol. 1. Morgan Kaufmann, Burlington (1993)
17. Reshma, U., Barathi Ganesh, H.B., Anand Kumar, M., Soman, K.P.: Supervised methods for domain classification of tamil documents. *ARPN J. Eng. Appl. Sci.* **10**(8), 3702–3707 (2015)
18. Seshadri, S., Madasamy, A.K., Padannayil, S.K., Anand Kumar, M.: Analyzing sentiment in indian languages micro text using recurrent neural network. *IIOAB J.* **7**, 313–318 (2016)