International Journal of Future Generation Communication and Networking
Vol. 13, No. 1s, (2020), pp. 257-- 264

Sentiment Analysis of Textual Data using various ML Techniques: A comparative study

Hemil Shah[1], Heyt Gala[2], Naman Shah[3], Ishani Saha[4]
Computer Engineering
[1,2,3,4]Mukesh Patel School of Technology Management & Engineering
Mumbai, India

Abstract

Sentiment analysis is a type of opinion mining which is used to determine the person's opinion, feelings, thoughts and judgement expressed on the text. Sentiment analysis main concern deals with classifying what the person expresses in its text and analyzing these text helps to know whether the person is angry, sad, happy etc. So, in this paper we have classified the text in 3 parts positive, negative and neutral. The positive text means the person is happy, or supporting a good cause etc., negative text means either the person is angry, sad, upset etc. and neutral text deals with the person giving facts or information about something. This paper deals with how we have used three models for classification of text naïve bayes, random forest and support vector machine.
Keywords*: Sentiment Analysis; Machine Learning;Naïve bayes; Support Vector Machine(SVM); Random Forest classifier; Lemmatization; StopWords .*

1. Introduction

Sentiment Analysis is becoming the growing need for large companies, businessman, tutors etc. The sentiment analysis helps to know what people feel about a subject and whether the feedback by the person is good or bad. One of the best reasons of using sentiment analysis is that it analyses large amount of data and humans are bound to errors so if the same thing is done by the machine it will both save time and resources. This paper deals with how we analyzed the text and classified based on the sentiments. There are four sections in this paper section 2 will be literature review, section 3 consist of methodologies and section 4 is the conclusion.

2. Literature Review

Sentiment Analysis is one of outgrowing fields which help many things which is useful for easier working process. With the use of sentiment analysis on twitter data, we were able to identify how the people felt about current scenarios happening all over the world. For e.g. The new law passed by the government of India for citizenship of Muslims, a large no. of people commented about it on twitter and we were able to identify the negative and positive comments given by the people.

As mentioned in paper [1] they classified tweets using the HMM model, SentiWordNet, Naïve Bayes and ensemble approach and giving as much 70% accuracy. In this paper, sentiwordnet classifies the tweet as positivity, negativity and objectivity. This method is used to tag the words used in the sentences. They have used the lexicon-based approach which means it maps the word to the dictionary meaning of words. The HMM model used is used to analyze sentiment tag and also see the occurrence of that word.

Another paper [2] did the same classification of twitter to analyze public reaction to UK energy companies. They have used the lexicon approach too but in different way. They have used lexicon approach for tagging words and also for classification using two methods Sentimentr and Hu & Liu Lexicon.

There have been many papers who have used different methods for classification, like in paper [3] they used aspect level approach to classify e-commerce data like public reactions. They used sentiwordnet for tagging the sentence whether it was positive, negative or neutral and then they classified the tweets using SVM and Naïve Bayes algorithm.

In paper [6] wherein the prediction of election result for Donald Trump and Hillary Clinton was analyzed for 6 weeks period wherein they used NRC classifier which is a lexicon approach and it

ISSN: 2233-7857 IJFGCN

consist of almost 31,000 words which can be used tagging the sentence and thus classifying the tweets using that.

The sentiment analysis is also used by institutes to analyze the feedback of the students so in the paper [7], they preprocessed the data and extracted features using n-grams and tf-idf and then mapped words to the lexicon features. Then they used the classification algorithms as SVM and Random Forest classifier and thus simplifying the analysis of the feedbacks.
So, sentiment analysis is not only helpful for companies but also for the government to know how the citizens feel about their decisions and how they could take their opinions into consideration. So, in this paper we have used sentiment analysis on twitter to classify negative and positive tweets.

3. Methodology

This section mainly deals with classification of the texts and sentiment classification. The sentiment classification is not any easy task it requires lot of preprocessing of data so that data could be easily classified. To achieve the result there are various modules.

1. Data collection/ data acquisition
2. Preprocessing
3. Feature Extraction
4. Sentiment Classification with methods and classifiers

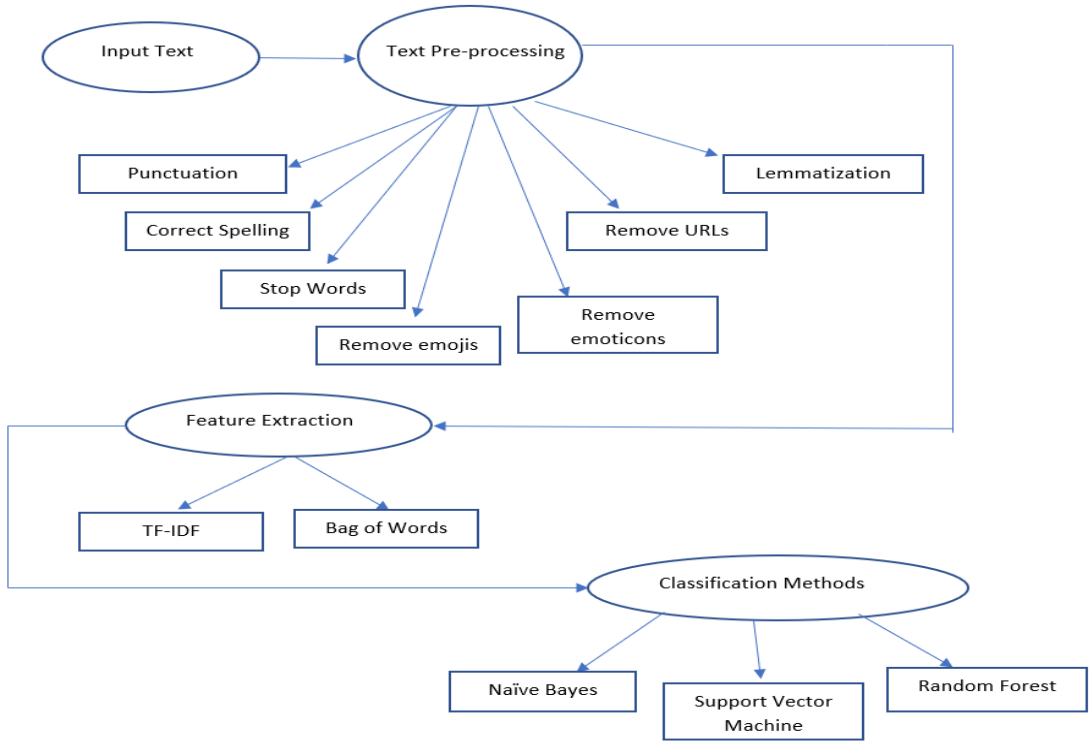


Figure 1. Process of classification of twitter data

3.1 Data Acquisition/ collection:

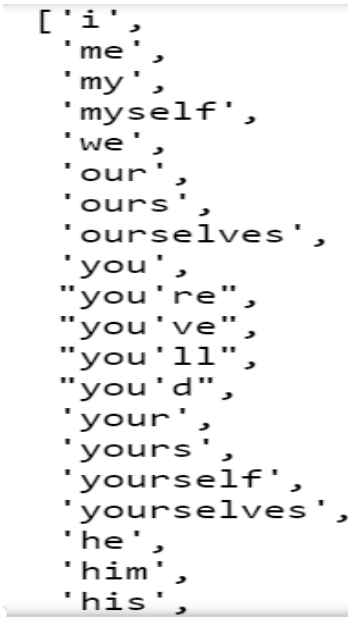
For training the data we used labeled datasets which is available on Kaggle. These datasets contained almost 50,000 tweets with labels of 0 and 1. 0 determining the negative behavior and 1 determining the positive behavior. So, we trained 70% of that data and tested 30% of the data.

3.2 Preprocessing:

In this process data is cleaned and it is presented like a natural sentence. The special symbols like" #, $, @" and also the URLs for any site are removed. In many of the sentences we found spelling

ISSN: 2233-7857 IJFGCN

errors, punctuations etc. so for that we used functions. We also used stopwords library which is has almost 29,000 words which are commonly used so we removed them. Also, all the emojis and emoticons are crucial part of tweet, so we marked smiley emoji as happy hilarious and removed the emojis after tagging them in terms of the text. After that we used tokenization method for processing the tweets. Then we used lemmatization method, which is used for tagging adverbs, nouns, adjectives etc. and getting them in single form. For eg: running, ran, run all comes down to 'run'.



```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his',
```

Figure 2. The above diagram are a few words which are used in preprocessing of the data and these are as many 29,000.

tweet	text_lower	text_punct	text_stop	text_common	text_rare	text_token	text_lemma
@user when a father is dysfunctional and is s...	@user when a father is dysfunctional and is s...	user when a father is dysfunctional and is so...	user father dysfunctional selfish drags kids d...	father dysfunctional selfish drags kids dysfun...	father dysfunctional selfish drags kids dysfun...	[father, dysfunctional, selfish, drags, kids, ...	father dysfunctional selfish drag kid dysfunct...
@user @user thanks for #lyft credit i can't us...	@user @user thanks for #lyft credit i can't us...	user user thanks for lyft credit i cant use ca...	user user thanks lyft credit cant use cause do...	thanks lyft credit cant use cause dont offer w...	thanks lyft credit cant use cause dont offer w...	[thanks, lyft, credit, cant, use, cause, dont,...	thanks lyft credit cant use cause dont offer w...
bihday your majesty	bihday your majesty	bihday your majesty	bihday majesty	bihday majesty	bihday majesty	[bihday, majesty]	bihday majesty
#model i love u take with u all the time in ...	#model i love u take with u all the time in ...	model i love u take with u all the time in u...	model love u take u time urð ðððð ðð	model take urð ðððð ðð	model take urð ðððð ðð	[model, take, urð, ðððð, ðð]	model take urð ðððð ðð
factsguide: society now #motivation	factsguide: society now #motivation	factsguide society now motivation	factsguide society motivation	factsguide society motivation	factsguide society motivation	[factsguide, society, motivation]	factsguide society motivation

Figure 3. Figure showing how the textual data is preprocessed column by column and final column has clean data

3.3 Feature Extraction:

ISSN: 2233-7857 IJFGCN

In feature extraction important words or emotion are extracted. For feature extraction we used two methods TF-IDF and bag of words (BOW).

TF-IDF:

TF (term frequency): this method is used to get the frequency of the words that are there in a particular sentence.

IDF (inverse document frequency): this method is used to see the frequency of the words in the entire dataset.

So, TF-IDF is basically used to get the rare words. The count of words in the document decides its rarity. For eg: 'the','is','have' etc. are the kind of words which are commonly used so TF-IDF basically transforms the sentence so that we get all the rare words.

Bag of Words:

Bag of Words is usually used to split the words in the sentence and thus getting the words on which TF-IDF is performed. The bag of words is also used for extracting the feature as it is also used to get the rare words which would mean getting a good feature out of the sentence.

3.4 Sentiment classification with methods and classifiers:

We will be classifying the text if it is positive or negative. After extracting the features, we would now train the model and test the model with various methods. We had split the datasets randomly in terms of (70,30). The 70% of the data is used for training the model and 30% is used for testing the model.

There are many methods and classifier used for classifying the sentiments.

Sentiment classification using Naïve Bayes:

Naïve Bayes is probability-based algorithm. It is used to compute the probability of every aspect of the text so in this case the features of the text. In Naïve Bayes we usually get the probability with aspect of the feature and then the total probability is bifurcated based on which we wanted to classify. In our case, it was whether the tweet was negative or positive.

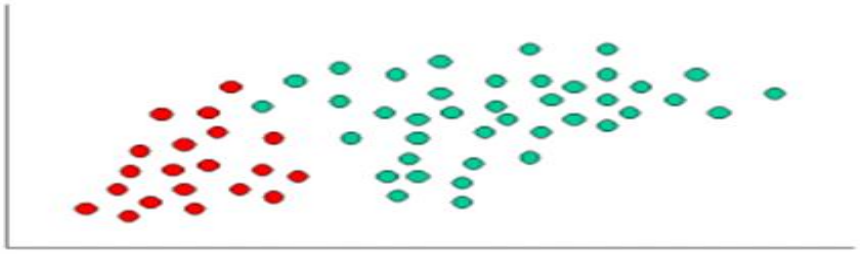


Figure 4. Naïve Bayes classification

For naïve bayes classification we received about 94% of accuracy and below is the figure shows that:

	precision	recall	f1-score	support
0	0.94	1.00	0.97	8905
1	1.00	0.15	0.26	684
micro avg	0.94	0.94	0.94	9589
macro avg	0.97	0.57	0.61	9589
weighted avg	0.94	0.94	0.92	9589

Figure 5. The above figure shows the precision, recall and F1-score of the classification of tweets using Naïve bayes .

ISSN: 2233-7857 IJFGCN

Sentiment classification using Random Forest Classifier:
Random Forest classifier is another method which analyses the tweets and classifies them. The random forest classifier is basically like a collection of decision trees and the result of a tweet in each decision tree are compared with each other and then the final output for the tweet is given. In our case, we used 200 decision trees in random forest to get an optimal output. The random forest proved to be great in classifying tweets.

Figure 6. Working of the Random Forest Classifier [5]

	precision	recall	f1-score	support
0	0.96	0.99	0.98	8905
1	0.85	0.48	0.62	684
micro avg	0.96	0.96	0.96	9589
macro avg	0.91	0.74	0.80	9589
weighted avg	0.95	0.96	0.95	9589

Figure 7. The above figure shows the precision, recall and F1-score of the classification of tweets using Random Forest Classifier.

Support Vector Machine:
It uses the hyper plane to divide the datasets into classes. Hyper plane classifies the set of data. Support Vector Machine is one of the classification techniques which is great at classifying the tweets. As we used SVM we had to modify and iterate many values in SVM. The gamma and C value iteration in SVM was done by us so that we get a good accuracy. There were about 90 iterations in C and gamma values done by us which then gave us the best fit value of C and gamma

ISSN: 2233-7857 IJFGCN

thus using those values for classification it proved to be highly effective helping us get 96% of accuracy in sentiment analysis.

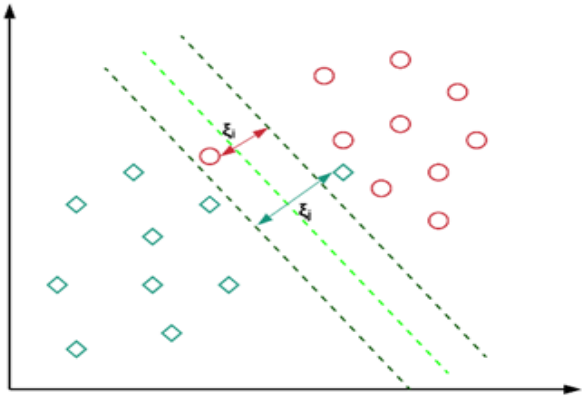


Figure 8. SVM classification
[4]

Below figure shows the iterations done on the C and gamma values of SVM:

```
Fitting 3 folds for each of 30 candidates, totalling 90 fits
[CV] C=0.001, gamma=5 ................................................
[CV] ....... C=0.001, gamma=5, score=0.9302855610671672, total=   3.6s
[CV] C=0.001, gamma=5 ................................................
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:    6.0s remaining:    0.0s
[CV] ....... C=0.001, gamma=5, score=0.9304009655357383, total=   3.7s
[CV] C=0.001, gamma=5 ................................................
[Parallel(n_jobs=1)]: Done   2 out of   2 | elapsed:   12.3s remaining:    0.0s
[CV] ....... C=0.001, gamma=5, score=0.9304009655357383, total=   3.6s
[CV] C=0.001, gamma=1 ................................................
[CV] ....... C=0.001, gamma=1, score=0.9302855610671672, total=   3.8s
```

Figure 9. The C and gamma value iterations

	precision	recall	f1-score	support
0	0.97	0.99	0.98	8905
1	0.84	0.56	0.67	684
micro avg	0.96	0.96	0.96	9589
macro avg	0.90	0.78	0.83	9589
weighted avg	0.96	0.96	0.96	9589

Figure 10. The above figure shows F1-score, recall and precision for classification of tweets using SVM.

RESULTS:

Table 1. The below table is the confusion matrix of 9,589 tweets of Naïve Bayes.

Classifier	True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)

ISSN: 2233-7857 IJFGCN

Naïve Bayes	8905	583	101	0

Table 2. The below table is the confusion matrix of 9,589 tweets of SVM classifier.

Classifier	True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)
SVM	8832	301	383	73

Table 3. The below table is the confusion matrix of 9,589 tweets of Random Forest classifier.

Classifier	True Positive (TP)	False Positive (FP)	True Negative (TN)	False Negative (FN)
Random Forest	8848	354	330	57

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 4: Comparing the F1-scores of all three methods

Classifier	**F1-Score (%)**
Naïve Bayes	94%
SVM	96%
Random Forest	96%

ISSN: 2233-7857 IJFGCN

4. Conclusion

We used three methods and they proved to be very effective and thus it can be used for many other applications. The future scope of sentiment analysis is very vast and thus it can help many different companies to get customers, help the government, for tutorial feedback analysis etc.

5. Future Work

As it has vast applications, we will use sentiment analysis for following purpose:

- Prediction of election results
- Analysis of student feedback
- Verbal Sentiment Analysis

References:

1. Rincy Jose and Varghese S Chooralil, "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Classifier EnsembleApproach," International Conference on Data Mining and Advanced Computing (SAPIENCE), Ernakulam, India, December 2016.

2. Victoria Ikoro, Maria Sharmina, Khaleel Malik and Riza Batista- Navarro, " Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers," Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS), Valencia, Spain, December 2018.

3. Satuluri Vanaja and Meena Belwal," Aspect-Level Sentiment Analysis on E-Commerce Data," International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, January 2019

4. https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe

5. https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/

6. Satish M. Srinivasan, Raghvinder S. Sangwan,Colin J. Neill, and Tianhai Zu," Twitter Data for Predicting Election Results: Insights from Emotion Classification," IEEE Technology and Society Magazine,Vol.38 , Issue. 1 , March 2019

7. Zarmeen Nasim, Quratulain Rajput and Sajjad Haider," Sentiment Analysis of Student Feedback Using Machine Learning and Lexicon Based Approaches," International Conference on Research and Innovation in Information Systems (ICRIIS), Langkawi, Malaysia, August 2017

ISSN: 2233-7857 IJFGCN