International Journal of Research and Development in Applied Science and Engineering (IJRDASE)
ISSN: 2454-6844

Opinion Mining- Twitter Data using Support Vector Machine Approach

Rudrendra Bahadur Singh
Department of CSE
Integral University, Lucknow
rudra.rathor20@gmail.com

Mohammad Arif
Department of CSE
Integral University, Lucknow
arif_mohd2k@yahoo.com

Mohd Akbar
Department of CSE
Integral University, Lucknow
akbar@iul.ac.in

***Abstract*--Due to the large amount of user opinions, reviews, comments, feedbacks and suggestions it is essential to explore, analyze and organize the content for efficient decision making. In the past years sentiment analysis has emerged as one of the popular techniques for information retrieval and web data analysis. In this paper we explored the machine learning classification approaches with different feature selection schemes to obtain a sentiment analysis model for the twitter dataset. Natural Language Processing (NLP) and Machine Learning approaches were used for the process. Multiple experiments were carried out using different feature sets and parameters to obtain maximum accuracy. Model implementation has been carried out using Support Vector Machine (SVM) learner. From the Results and Analysis it is seen that the performance of this SVM Model has a better accuracy of 81.50%, with the recall of both positive and negative sentiments being more or less same, each being 83% and 80% respectively. Also the model performance has RMSE value of 0.395 and Absolute error of 0.369, which again clearly demonstrates the good predictive capability of the model.**

Key Words-- Sentiment Analysis, SVM, RMSE, NLP

1 Introduction

The evolution of web technology has led to a huge amount of user generated content and has significantly changed the way we manage, organize and interact with information. Due to the large amount of user opinions, reviews, comments, feedbacks and suggestions it is essential to explore, analyze and organize the content for efficient decision making. In the past years sentiment analysis has emerged as one of the popular techniques for information retrieval and web data analysis. Sentiment analysis, also known as opinion mining is a subfield of Natural Language Processing (NLP) and Computational Linguistics (CL) that defines the area that studies and analyzes people's opinions, reviews and sentiments.

Sentiment analysis defines a process of extracting, identifying, analyzing and characterizing the sentiments or opinions in the form of textual information using machine learning, NLP or statistics. A basic sentiment analysis system performs three major tasks for a given document. Firstly it identifies the sentiment expressing part in the document. Secondly, it identifies the sentiment holder and the entity about which the sentiment is expressed. Finally, it identifies the polarity (semantic orientation) of the sentiments.

Sentiment analysis can be performed at three different levels: document, sentence and aspect level. The document level sentiment analysis aims at classifying the entire document as positive or negative, (Pang et al, [1]; Turney, [2]). The sentence level sentiment analysis is closely related to subjectivity analysis. At this level each sentence is analyzed and its opinion is determined as positive, negative or neutral, (Riloff et al, [3]; Terveen et al, [4]). The aspect level sentiment analysis aims at identifying the target of the opinion. The basis of this approach is that every opinion has a target and an opinion without a target is of limited use [5]. In this paper we explored the machine learning classification approaches with different feature selection schemes to obtain a sentiment analysis model for the twitter dataset. Natural Language Processing and Machine Learning approaches were used for the process. Multiple experiments were carried out using different feature sets and parameters to obtain maximum accuracy.

2 Data Used

The proposed work is evaluated by running experiments with the twitter dataset, available at http://help.sentiment140.com/for-students/ . Sentiment model has been built using supervised learning. For this a set of 200 twitter data available from corpus data found at: http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip has been used. It has 200 positive reviews, 200 negative reviews and 200 unlabelled reviews for testing of the model.

3 Dataset Pre-Processing

The initial Sentiment 200 dataset, downloaded from the internet is subjected to pre-processing so as to be able to be used for sentiment analysis using Rapid Miner. Initially the dataset so downloaded were in .csv format in Excel sheet, having positive, negative and neutral tweets along with tweet id, date, query, etc. All these were filtered except for sentiments and tweets. Next the dataset is subjected to cleaning. The aim of the data cleaning process is to remove any unwanted content from the training data and the input tweets. The term unwanted content is used to describe any piece of information within the tweet that will not be useful for the machine learning algorithm to assign a class to that tweet. Data cleaning can not only simplify the classification task for the machine learning model but it also serves to greatly decrease processing cost in the training phase. The unwanted content is tabulated in Table 1 below.

Table 1: Unwanted Contents Removed from the data

Unwanted Content	ACTION
Punctuation (! ? , . " : ;)	Removed
#word	Removed #
@user	Replaced with "AT_USER"
RT	Removed
Emoticons	Removed
Unicode formatting	Removed
Uppercase characters	Lowercase all content
URLs and web links	Replaced with "URL"

Once the data has been cleaned, it is again saved in a separate excel sheets having positive and negative tweets only. The tweets stored in diferrent columns, both for positive and negative datasheets are further converted into separate wordpad files containing each tweets. Thus after final processing there were 200 notepads each for negative and positive sentiments saved in separate folders. Also a mixed tweets consisting of both negative and positive ones were saved in a separate folder named as unlabelled folder.

4. Problem Statement and Proposed Technique

This section presents the proposed technique to analyze sentiments in a twitter domain. The proposed approach uses a combination of NLP techniques and supervised learning. In the first stage a pre-processing model is proposed to optimize the dataset. In the second stage experiments are performed using the machine learning methods to obtain the performance vector for various feature selection schemes. We used up to 4-grams (i.e. n=1, 2, 3, 4) in this work. First step for implementing this analysis is Pre-Processing the document from data i.e. extracting the positive and negative reviews of twitter and storing it in different polarity. At first, both positive and negative reviews are taken. All of the words are stemmed into root words. Then the words are stored in different polarity (positive and negative). Both vector wordlist and model are created. Next, the required list of unlabelled data is given as input to the model validation. Model compares each and every word from the given list of data with that of words which come under different polarity stored earlier. The review is estimated based on the majority of number of words that occur under a polarity.

The flowchart used in Rapid Miner for carrying out sentiment analysis are given in Figures 1, 2 and 3.

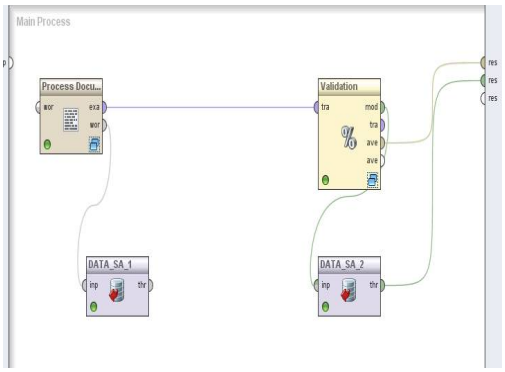


Fig. 1. Initial Stages for Sentiment Analysis

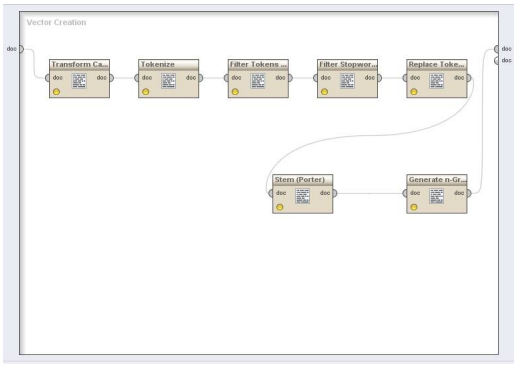


Fig. 2. Stages for Processing of Documents

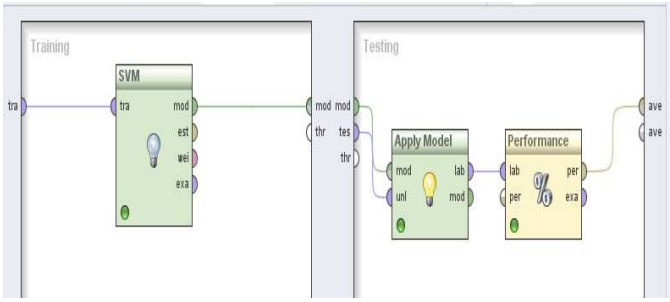


Fig. 3. Training and Testing phase for the Validation Model

5. Parameters Used:

Table 2 : Parameter values of the Operators

Sl. No.	Operator		Parameter Used	Type
1	**Process Document**	**a**	Word vector creation scheme	TF-IDF
		b	Prune Method	percentual
				(<3% and >95%)
2	Transform cases			lower case
3	Tokenize			non letters
4	Filter Tokens	a	min. Char.	4
		b	max. Char.	25
5	Validation	a	no. of validations	10
		b	sampling type	Automatic
6	SVM	a	Kernel type	**dot**
		b	Kernel cache	200
		c	Convergence epsilon	0.001
		d	Maximum iterations	100000

7	Performanc e		Accuracy	
			RMSE	
			Absolute error	

6 Result Analysis:

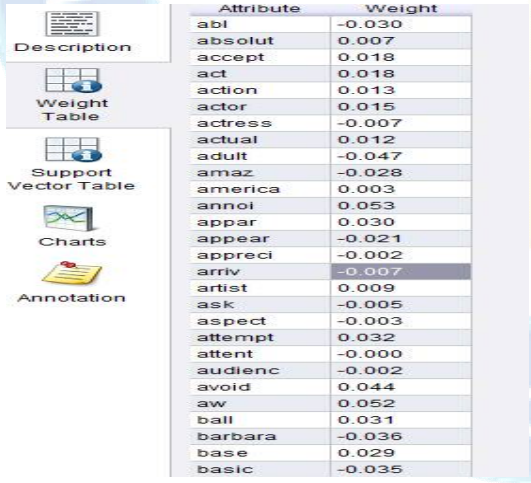
The dataset consists of 400 reviews equally divided into 200 positive and 200 negative. The dataset used for the experiments was divided into two classes, positive and negative. For a given classifier and a document there are four possible outcomes: true positive, false positive, true negative and false negative. If the document is labelled positive and is classified as positive it is counted as true positive else if it is classified as negative it is counted false negative. Similarly, if a document is labelled negative and is classified as negative it is counted as true negative else if it is classified as positive it is counted as false positive. Based on these outcomes a two by two confusion matrix can be drawn for a given test set. This is shown in Figure 8 below.

The confusion matrix in figure 8 forms the basis for the calculation of the following metrics.

i. *Accuracy* = (tp+tn)/ (P+N)
ii. *Precision* = tp/ (tp+fp)
iii. *Recall/ true positive rate* = tp/P
iv. *F-measure* =2/ ((1/precision)+(1/recall))
v. *False alarm rate/ false positive rate* = fn/N
vi. *Specificity* = tn/ (fp+tn) = (1-fp rate)

The experiments show that Term frequency-Inverse document frequency (TF-IDF) scheme gives maximum accuracy using SVM classification approach.

From the SVM model built from our dataset with sentiment attributes, the various attributes with weights defined are generated, a snapshot of it given in Fig 4 below, with the kernel model of various attributes given in Fig. 5 below.



Attribute	Weight
abl	-0.030
absolut	0.007
accept	0.018
act	0.018
action	0.013
actor	0.015
actress	-0.007
actual	0.012
adult	-0.047
amaz	-0.028
america	0.003
annoi	0.053
appar	0.030
appear	-0.021
appreci	-0.002
arriv	-0.007
artist	0.009
ask	-0.005
aspect	-0.003
attempt	0.032
attent	-0.000
audienc	-0.002
avoid	0.044
aw	0.052
ball	0.031
barbara	-0.036
base	0.029
basic	-0.035

Fig. 4. Snapshot of Weights of Attributes generated by SVM Model

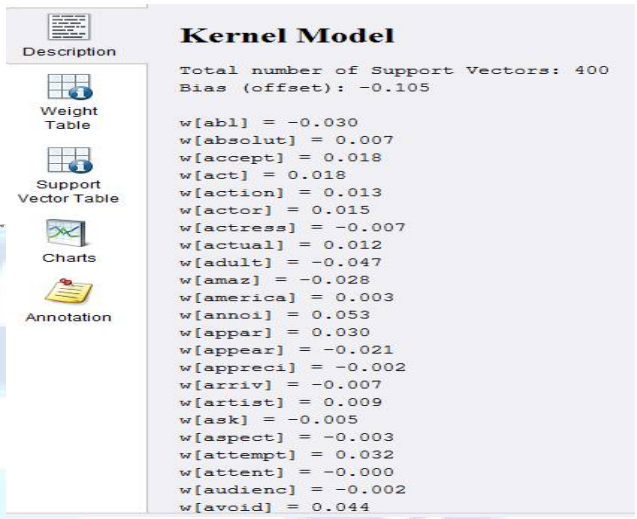


Fig 5. Snapshot of the SVM Kernel Model

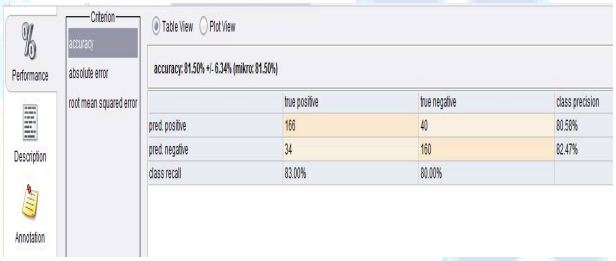


Fig. 6. Snapshot of the Model Performance using SVM Classification approach

From the Performance operator, we get various measures of the sentiment dataset, as seen from Figure 6 above.

The performance of this SVM Model has a better accuracy of 81.50% (Fig. 6 & 7), with the recall of both positive and negative sentiments being more or less same, each being 83% and 80% respectively. Also the model performance has RMSE value of 0.395 and Absolute error of 0.369, which again clearly demonstrates the good predictive capability of the model.

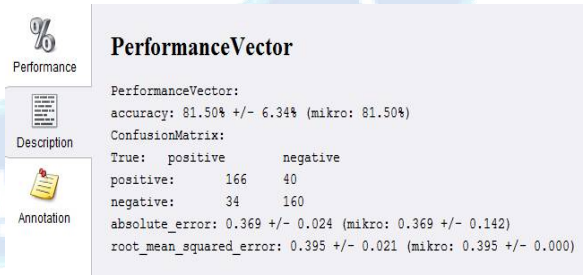


Fig. 7. Snapshot of the performance of the model

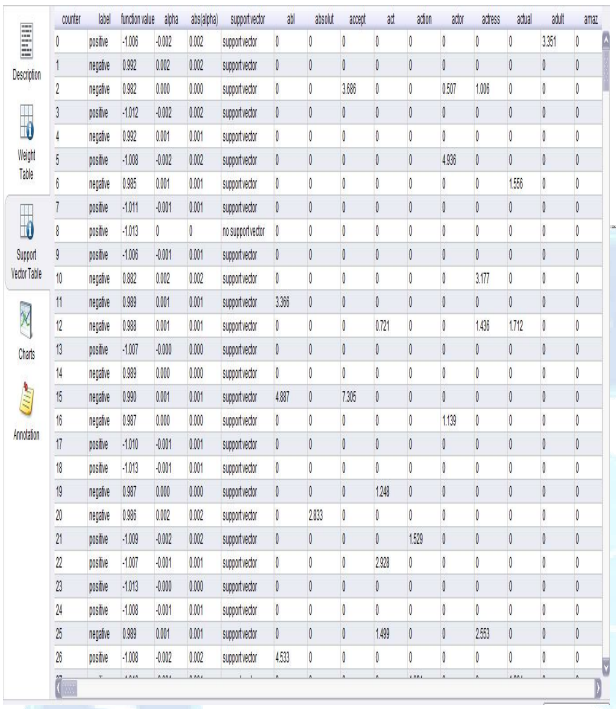


Fig 8. Snapshot of the SVM analysis

Fig. 8 above and Fig. 9 and 10 below are the snapshots of the SVM analysis of the unlabelled data subjected to the SVM model developed before. From Fig. 9 given below, it can be seen that of the 200 unlabelled datasets fed into the model, the model was successfully able to predict the sentiments as 78 negative and 122 positive, with average positive confidence of 0.534 and negative confidence of 0.466, which again is a good predictability indicator. The detailed analysis of the unlabelled are shown in Fig. 10 below.

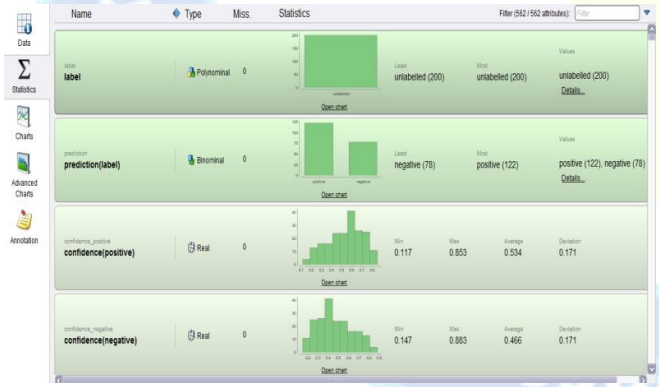


Fig. 9: of the statistical view of the Prediction results of unlabelled data

7. Conclusion

Here machine learning classification approaches with different feature selection schemes to obtain a sentiment analysis model for the twitter dataset has been explored. Natural Language Processing and Machine Learning approaches were used for the process. Multiple experiments were carried out using different feature sets and parameters to obtain maximum accuracy.

Fig. 10. Snapshot of the confidence of unlabelled data derived from the model

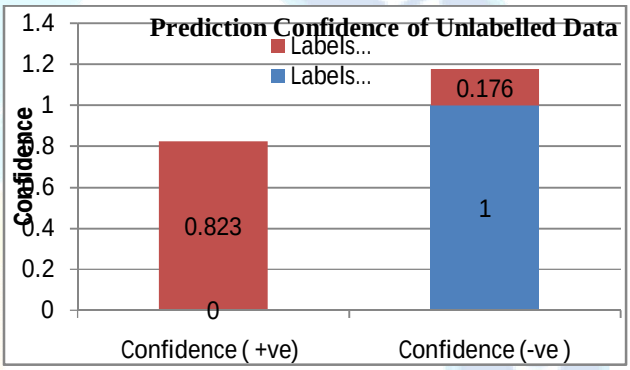


Fig. 11: Prediction results of Unlabelled Data

The proposed work is evaluated by running experiments with the twitter dataset. The dataset consists of 400 reviews equally divided into 200 positive and 200 negative. The proposed approach uses a combination of NLP techniques and supervised learning. In the first stage a pre-processing model is proposed to optimize the dataset. In the second stage experiments are performed using the machine learning methods to obtain the performance vector for various feature selection schemes.

Here Rapid Miner Studio 6.5 software with the text processing extension, licensed under AGPL version3, and Java1.6 has been used. Rapid Miner supports the design and documentation of overall data mining process. Model implementation has been carried out using Support Vector Machine (SVM) learner. From the Results and Analysis it is seen that the performance of this SVM Model has a better accuracy of 81.50%, with the recall of both positive and negative sentiments being more or less same, each being 83% and 80% respectively. Also the model performance has RMSE value of 0.395 and Absolute error of 0.369, which again clearly demonstrates the good predictive capability of the model. Further when the model so generated was subjected to unlabelled data analysis, it is seen that the average confidence (positive) and confidence (negative) are 0.534 and 0.466 respectively, which again clearly

demonstrated the good sentiment analysis of the unlabelled data using SVM classifier.

References:

[1] B. Pang et al, 2002, Thumbs up ?: sentiment classification using machine learning techniques, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, vol.10, 79-86.
[2] P.D. Turney, 2002, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the Association for Computational Linguistics (ACL), 417–424.
[3] Riloff, E &Wiebe, J., 2003, Learning extraction patterns for subjective expressions, EMNLP"03.
[4] Loren Terveen et al, 1997, PHOAKS: A system for sharing recommendations, Communications of the Association for Computing Machinery (CACM), 40(3):59–62.
[5] Minqing Hu and Bing Liu, 2004, Mining and summarizing customer reviews, Proceedings of the 10th ACM SIGKDD International conference on knowledge discovery and data mining.