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

Sentiment Analysis of Twitter Data Using Classification Approach

Saumya Shukla
Computer Science & Engg. Department
AIET, Lucknow, India
saumyashukla2307@gmail.com

Manmohan Singh Yadav
Computer Science & Engg. Department
AIET, Lucknow, India
yadavmohanman100@gmail.com

Abstract--The evolution of web technology has led to a huge amount of user generated content and has significantly changed the way one manages, organizes and interacts with information. sentiment analysis has emerged as one of the popular techniques for information retrieval and web data analysis. In the present paper machine learning classification approaches with different feature selection schemes to obtain a sentiment analysis model for the twitter dataset has been explored. The proposed approach uses a combination of Natural Language Processing (NLP) techniques and supervised learning Support Vector Machine (SVM) classifier. The work has been performed using Rapid Miner tool. 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.

Keywords: Sentiment Analysis, NLP, SVM, RMSE, Rapid Miner.

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. 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.

2. Related Work

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, Amit P. Sheth, showed that user generated content on Twitter (produced at an monumental rate of 340 million tweets per day) provides a rich supply for gleaning people's emotions, which is necessary for deeper understanding of people's behaviors and actions. **Soujanya Poria, Erik Cambria, Alexander Gelbukh, Federica Bisio, Amir Hussain** showed how computational intelligence techniques are combined with common-sense computing and linguistics to analyze sentiment data flows, i.e., to automatically decrypt however humans categorical emotions and opinions via natural language. **Otto K. M. Cheng, Raymond Lau ,** showed the illustration of the event of a completely unique big data stream analytics framework named BDSASA that leverages a probabilistic language model to investigate client the buyer the patron sentiments embedded in many countless on-line consumer reviews. **Bingwei Liu, Erik Blaschy, Yu Chenz, Dan Shen_ and Genshe Chen** aimed to value the measurability of Na¨ıve bayes classifier (NBC) in massive datasets. Instead of employing a standard library (e.g., Mahout), authors implemented NBC to come through fine-grain management of the analysis procedure. **Ahmad Ghazal1,Francois Raab, Meikel Poess, Alain Crolotte, Hans-Arno Jacobsen** in their work presented BigBench, an end-to-end big data benchmark proposal. The underlying business model of BigBench is a product retailer. The proposal covers a data model and synthetic data generator that addresses the range, velocity and volume aspects of big data systems con-taining structured, semi-structured and unstructured data. **Basant Agarwal, Namita Mittal, Pooja Bansal, and Sonal Garg** proposed a novel sentiment analysis model supported common-sense information extracted from construct internet based mostly ontology and context data. Concept internet based mostly ontology is used to see the domain specific ideas that successively created the domain specific vital options. **Marcos D. Assun, Rodrigo N. Calheirosb, Silvia Bianchic, Marco A. S. Nettoc, Rajkumar Buyyab** discussed approaches and environments for carrying out analytics on Clouds for big Data applications. **Yang Yu, Xiao Wang,** used sentiment analysis to examine U.S. soccer fans' emotional responses in their tweets, particularly, the emotional changes after goals (either own or the opponent's). Authors found that throughout the matches that the U.S. team played, fear and anger were the most common negative emotions and generally, increased once the opponent team scored and bated once the U.S. team scored.

3. 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.

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. Rapid Miner Studio 6.0 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.

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.

A. Stages in Sentiment Analysis of Twitter Data

A basic task in sentiment analysis is classifying an expressed opinion in a document, a sentence or an entity feature as positive or negative. The work here gives the list of twitter data and its review such as **Positive** or **Negative**. This program implements Precision and Recall method. **Precision** is the probability that a (randomly selected) retrieved document is relevant. **Recall** is the probability that a (randomly selected) relevant document is retrieved in a search. Or high **recall** means that an algorithm returned most of the relevant results. High **precision** means that an algorithm returned more relevant results than irrelevant.

B. Parameters Used

Table 1 : 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	Performance		Accuracy	
			RMSE	
			Absolute error	

5. 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 1 below.

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

i. *Accuracy* = (tp+tn)/ (P+N)

ii. *Precision* = tp/ (tp+fp)

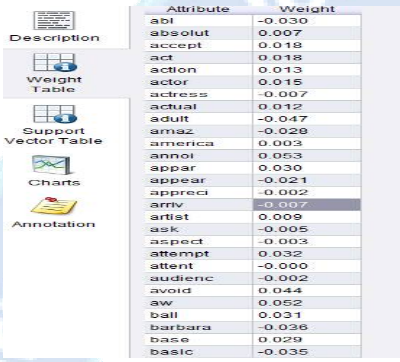
iii. *Recall/ true positive rate* = tp/P

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

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)

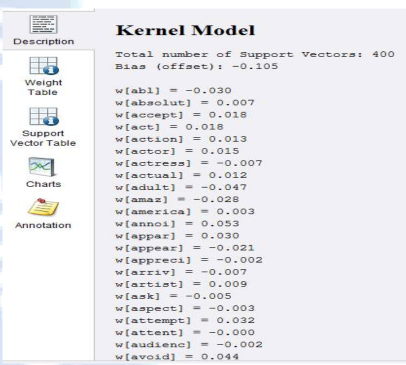
The experiments show that Term frequency-Inverse document frequency (TF-IDF) scheme gives maximum accuracy using SVM classification approach.

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 1 below, with the kernel model of various attributes given in Fig. 2 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. 1: Snapshot of Weights of Attributes generated by SVM Model




```
Kernel Model

Total number of Support Vectors: 400
Bias (offset): -0.105

w[abl] = -0.030
w[absolut] = 0.007
w[accept] = 0.018
w[act] = 0.018
w[action] = 0.013
w[actor] = 0.015
w[actress] = -0.007
w[actual] = 0.012
w[adult] = -0.047
w[amaz] = -0.028
w[america] = 0.003
w[annoi] = 0.053
w[appar] = 0.030
w[appear] = -0.021
w[appreci] = -0.002
w[arriv] = -0.007
w[artist] = 0.009
w[ask] = -0.005
w[aspect] = -0.003
w[attempt] = 0.032
w[attent] = -0.000
w[audienc] = -0.002
w[avoid] = 0.044
```

Fig 2: Snapshot of the SVM Kernel Model

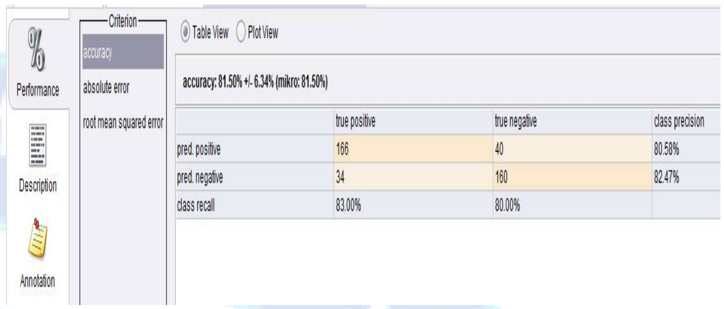
From the Performance operator, one get various measures of the sentiment dataset, as seen from Figure 3 below. The various performance measures are as mentioned below.

Accuracy is calculated by the percentage of correct predictions over the total number of examples. Correct prediction means examples where the value of the prediction attribute is equal to the value of the label attribute.

Precision of a class is calculated by taking the correct predictions of a label's value over the total predictions for the same label value (correct predictions + wrong predictions).

Recall of a class is calculated by taking the correct predictions of a label's value over the total of the real examples with the same label value (correct predictions + missed examples).

The performance of this SVM Model has a better accuracy of 81.50% (Fig. 3), 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.



	true positive	true negative	class precision
pred. positive	166	40	80.58%
pred. negative	34	160	82.47%
class recall	83.00%	80.00%	

Fig 3: Snapshot of the Model Performance using SVM Classification approach

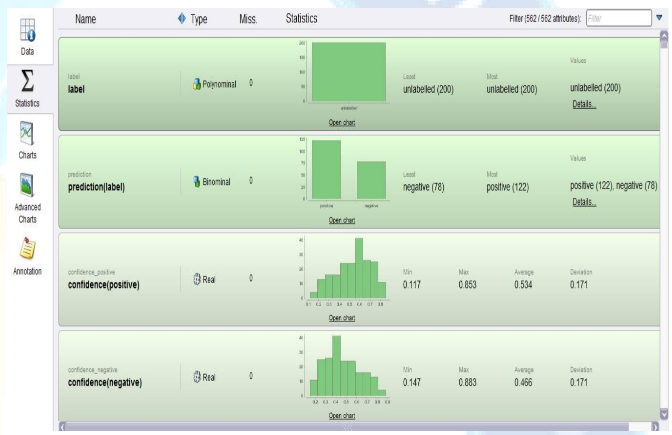
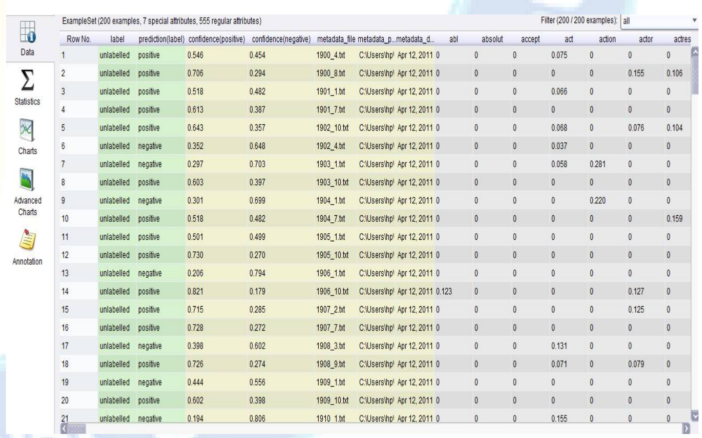


Fig 4: Snapshot of the statistical view of the Prediction results of unlabelled data



Row No.	label	prediction(label)	confidence(positive)	confidence(negative)	metadata_file	metadata_p...	metadata_d...	abl	absolut	accept	act	action	actor	actres
1	unlabelled	positive	0.546	0.454	1900_4.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0.075	0	0	0
2	unlabelled	positive	0.706	0.294	1900_8.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0.155	0.106
3	unlabelled	positive	0.518	0.482	1901_1.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0.066	0	0	0
4	unlabelled	positive	0.613	0.387	1901_7.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0	0
5	unlabelled	positive	0.643	0.357	1902_10.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0.068	0	0.076	0.104
6	unlabelled	negative	0.352	0.648	1902_4.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0.037	0	0	0
7	unlabelled	negative	0.297	0.703	1903_1.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0.058	0.281	0	0
8	unlabelled	positive	0.603	0.397	1903_10.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0	0
9	unlabelled	negative	0.301	0.699	1904_1.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0.220	0	0
10	unlabelled	positive	0.518	0.482	1904_7.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0	0.159
11	unlabelled	positive	0.501	0.499	1905_1.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0	0
12	unlabelled	positive	0.730	0.270	1905_10.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0	0
13	unlabelled	negative	0.206	0.794	1906_1.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0	0
14	unlabelled	positive	0.821	0.179	1906_10.txt	C:\Users\hp\	Apr 12, 2011	0.123	0	0	0	0	0.127	0
15	unlabelled	positive	0.715	0.285	1907_2.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0.125	0
16	unlabelled	positive	0.728	0.272	1907_7.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0	0
17	unlabelled	negative	0.398	0.602	1908_3.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0.131	0	0	0
18	unlabelled	positive	0.726	0.274	1908_9.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0.071	0	0.079	0
19	unlabelled	negative	0.444	0.556	1909_1.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0	0
20	unlabelled	positive	0.602	0.398	1909_10.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0	0	0	0
21	unlabelled	negative	0.194	0.806	1910_1.txt	C:\Users\hp\	Apr 12, 2011	0	0	0	0.155	0	0	0

Fig 5: Snapshot of the confidence of unlabelled data derived from the model

Further from Fig. 4 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. 5 above.

6. 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. 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]. Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, Amit P. Sheth, “Harnessing Twitter ‘Big Data’ for Automatic Emotion Identification”, http://blog.twitter.com/2012/03/twitter-turns-six.html

[2]. Soujanya Poria, Erik Cambria, Alexander Gelbukh, Federica Bisio, Amir Hussain, ARTICLE in IEEE COMPUTATIONAL INTELLIGENCE MAGAZINE, October, 2015Otto K. M. Cheng, Raymond Lau, (2015), “Big Data Stream Analytics for Near Real-Time Sentiment Analysis”, Journal of Computer and Communications, 3, 189-195 Published Online in SciRes. http://www.scirp.org/journal/jcc http://dx.doi.org/10.4236/jcc.2015.35024

[3]. Bingwei Liu, Erik Blaschy, Yu Chenz, Dan Shen_ and Genshe Chen, (2013), “Scalable Sentiment Classification for Big Data Analysis Using Naıve Bayes Classifier”, Big Data, IEEE International Conference, pp. 99-104.

[4]. Ahmad Ghazal1,Francois Raab, Meikel Poess, Alain Crolotte, Hans-Arno Jacobsen, (2013), “BigBench: Towards an Industry Standard Benchmark for Big Data Analytics”, *SIGMOD’13*, Copyright 2013 ACM 978-1-4503-2037-5/13/06

[5]. Basant Agarwal, Namita Mittal, Pooja Bansal, and Sonal Garg, (2015), “Sentiment Analysis Using Common-Sense and Context Information**”,** Computational Intelligence and Neuroscience, Volume 2015, Article ID 715730, 9 pages http://dx.doi.org/10.1155/2015/715730

[6]. Yang Yu, Xiao Wang, (20158), “World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans’ tweets”, Elsevier Computers in Human Behavior 48 (2015) 392–400

[7]. Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. Target dependent twitter sentiment classification. in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011). 2011.

[8]. Jindal, Nitin and Bing Liu. Identifying comparative sentences in text documents. in Proceedings of ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR-2006). 2006a.

[9]. Jindal, Nitin and Bing Liu. Mining comparative sentences and relations. In Proceedings of National Conf. on Artificial Intelligence (AAAI-2006). 2006b.

[10]. 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.

[11]. 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.

[12]. Riloff, E &Wiebe, J., 2003, Learning extraction patterns for subjective expressions, EMNLP‟03.

[13]. Loren Terveen et al, 1997, PHOAKS: A system for sharing recommendations, Communications of the Association for Computing Machinery (CACM), 40(3):59–62.

[14]. 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.