Sentiment Analysis Of Twitter Data

K S Kushwanth Ram[1], Sachin Araballi[2],Shambhavi B R[3],Shobha G[4]

Abstract— **Twitter is a popular micro blogging service where users create status messages or small text-based Web posts called tweets. Twitter currently receives in excess of340 million tweets a day, in which people share their comments regarding a wide range of topics. A large number of tweets include opinions about numerous subjects. Analyzing these tweets to extract opinions or sentiments help us determine the popularity of the subjects. This paper talks about a sentiment analyzer engine that can be used to analyze tweets. Tweets retrieved real time are classified as to belonging to one of positive, negative or neutral category using pre classified tweets as training data. The paper discusses about Naïve Bayes algorithm for implementing the sentiment analyzer engine. The Sentiment analyzer engine developed can give an approximate estimation of the success or popularity of a subject. The algorithm's efficiency is mainly dependent on the quality of the training data, for the training data chosen for this project we obtained an accuracy of close to 42% with precision and recall standing out at 45.65% and 67.74% respectively.**

Index Terms— **Naive Bayes Algorithm, Sentiment Analysis, Twitter, .**

I.INTRODUCTION

In the last couple of years the social medium Twitter has become more and more popular, Twitter is the most used micro blogging website with about 271 million active users generatingin excess of 340 million tweets a day; it is an interesting source of information. Tweets are a way to share interests publicly or among a designated private group. Twitter distinguishes itself from other social media by the limited message size. The maximum size of 140 characters restricts users in their writing. Twitter is therefore challenging their users to express their view in one or two key sentences.

Because Twitter is widely adopted through all strata, it can be seen as a good reaction of what is happening around the world. Among all that happens, the latest trends are most interesting for companies. The latest trends can be analyzed and when intended, reacted to. From a marketing point of view, these latest trends can be used to respond with appropriate activities, like product advertisements. Analyzing tweets can therefore be a goldmine for companies to create an advantage over competitors.

One interesting group is tweets expressing sentiments about products, brands or services. These messages contain an opinion about a specific subject. The sentiment of this opinion can be classified in different categories. An obvious example of three categories is positive, negative and neutral.

Algorithms have been developed which can be used to analyze data, with the goal to extract useful information. Some widely used classification algorithms from the literature are Naive Bayes and Support Vector Machines. This paper discusses about Naïve Bayes algorithm and the way it can be used to assign sentiment (positive, neutral or negative) to a tweet.

The objectives of the paper are 1) To give a detailed explanation of the Naïve Bayes algorithm, 2) To explain how tweets are analyzed using the Naïve Bayes algorithm 3) To compare the accuracy of using Naïve Bayes algorithm with that of other algorithms.

The paper is organised as follows, Section II discusses the various definitions and the meanings of terms.Section III outlines the challenges encountered for sentiment analysis. Section IV discusses the list of algorithms available for performing sentiment analysis. Section V deals with how sentiments of tweets were analysed using Naïve Bayes algorithm. Section VI talks about the results and accuracy of the sentiment analysis carried out using Naïve Bayes algorithm. Section VII compares accuracy of various other algorithms implemented for text classification. Section VIII is conclusion which talks about outcome of the work carried out and section IX talks about possible future enhancements.

ISSN: 2278 – 1323

II.DEFINITIONS

A. SENTIMENT ANALYSIS

- According to Wilson, Weber and Homann, "Sentiment analysis is the task of identifying positive and negative opinions, emotions, and evaluations".
- According to Liu [1] "Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text."

B.SENTIMENT

- According to WordNet a sentiment is "A personal belief or judgment that is not founded on proof or certainty"
- According to Wikipedia a sentiment is "An opinion is a subjective statement or thought about an issue or topic, and is the result of emotion or interpretation of facts."

III.CHALLENGES

A. SENTIMENT ANALYSIS AND CHALLENGES ASSOCIATED

Sentiment analysis involves several research fields: - natural language processing, computational linguistics and text analysis. It refers to the extraction of subjective information from raw data, often in text form. However other media types could contain subjective data, like images, sounds and videos but these types are less studied. In accordance, in all media types different kinds of sentiments exist. The sentiment can refer to opinions or emotions, even though these two types are related there is an evident difference. In sentiment analysis based on opinions, a distinction is made between positive, negative and neutral opinions. The sentiment analysis that is considered in this paper is based on opinions and is often referred in literature as opinion mining. Sentiment analysis aims to determine the attitude of the opinion holder with respect to a subject. Other applications try to determine the overall sentiment of a document. Sentiment analysis can be difficult. For example, a text can contain more than one opinion about the same object or about several objects.Opinion = $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$ Where o_j is particular object, f_{jk} is feature k of object o_j, h_i is an opinion holder, t_l is the time and oo_{ijkl} the actual opinion. Determining the actual polarity of some sentence is the most difficult task of the five properties of quintuple [2]. This sentiment is subjective because different people have different mental scale for what they consider to be a strong or a weak opinion. Therefore it can occur that somebody else can label a sentence that is labeled as positive by somebody as neutral.

B. TWITTER AND TWEETS

Twitter is a micro blogging website where users share information in the form of tweets. The information contained in the tweets have a maximum length of 140 characters. This limited number causes creative people to use acronyms and abbreviations to enlarge the expressibility of their message. Those acronyms lead to a broader dictionary of words, but also make it harder to analyze the tweets, since they create a broader feature space.

Another Twitter term is the retweet (RT), which is used to show the content of a tweet posted by another user. Users post retweets to note that the original message is interesting enough to send to their followers. An interesting question is whether one should include retweets for sentiment analysis, since it is actually a repetition of a tweet.

In the line of this research, emoticons are interesting, because they state the mood of a user. This mood is in some cases related and relevant for the sentiment of the message. Smiling and sad emoticons give a good indication of the sentiment; however other emoticons like confused or embarrassed are less informative. Therefore only a part of the emoticons could be useful for sentiment classification.

IV. ALGORITHMS

A. Naïve Bayes

Naïve Bayes Classifier is a probabilistic classifier based on applying Bayes' theorem with strong independence assumption that the presence of one feature in a class does not depend on the presence or absence of another feature.

The features or also known as attributes, are the characterized values to describe an instance. Individual instance is defined by its value on a fixed, predefined set of features or attributes [5]. For example, in the text classification problem, the features can be extracted from words in a document.

The independence assumption does not hold in real texts because of the grammatical relation between words in the sentence.

ISSN: 2278 – 1323

Naive Bayes [3] is a simple model which works well on text categorization. For tweets we use a multinomial Naive Bayes model. Class c* is assigned to tweet d, where $c* = argmax_c P_{NB}(c|d)$……..…(1.1)

and $P_{NB}(c|d) := \frac{P(c)\sum_{i=1}^{m} P(f|c)^{n_i(b)}}{P(d)}$ ……..(1.2)

In the equation (1.2), f represents a feature and $n_i(d)$ represents the count of feature f_i found in tweet d. There are a total of m features. Parameters P(c) and P (f_{jc}) are obtained through maximum likelihood estimates, and Laplace add-1 smoothing feature is utilized for unseen features [3].

With this algorithm, there are two representations of a document:

- Naïve Bayes Binary Model (NBB): only presence or absence of words is considered [4].
- Naïve Bayes Multinomial Model (NBM): multiple occurrences of words are considered [4].

For instance, the sentence "my brother is a teacher and my sister is a doctor" is represented as vector of words in two models as below:

- NBB: (my, brother, is, a, teacher, and, sister, doctor)
- NBM: (my, brother, is, a, teacher, and, my, sister, is, a, doctor)

B. Support Vector Machines

Another algorithm for solving the text classification problem is Support Vector Machine (SVM) introduced by [6]. The idea of this algorithm is to consider each document as a point in the document space and to find the appropriate hyper plane to separate the documents into two classes. Figure 1 depicts a sample view of the algorithm and documents belong to two classes and the hyper plane which separates them. The x and y are the co-ordinates of two dimensional space [6].

However, text classification problem involves with not only two classes but also multiple classes. So the algorithm needsto be extended. There are several works done with the extension of SVM [7]. Two simple approaches are:

- One against all: assume that there are only two classes, one class v/s other classes
- Pair wise classification: one class against one other class and aggregate the results

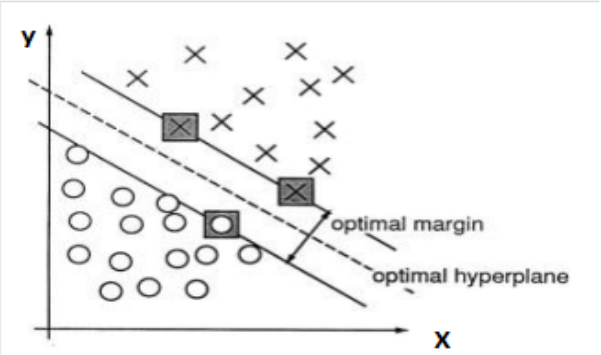

Figure 1 Support Vector Machine Illustration

V. IMPLEMENTATION

Tweets were retrived real time based on user query. The tweets were retrieved using the Twitter4j api's. These api's accepted a search term(usually a line of text or phrase) and returned the list of tweets that contained the search term. The tweets retrieved were later subjected to preliminary text processing, such as

- Replacing the URL's (Uniform Resource Identifier)present in the tweets with the key word URL.
- Replacing words directed towards a person like (@PersonName) with the word USERNAME.
- Removal of slang words from the tweets.
- Removal of articlesfrom the tweets.
- Removal of common adjectives from the tweets.

The above operations were done because all the above words or phrases do not have any sentiment associated with them.

A large number of tweets pre-classified as to be belonging to one of positive, negative and neutral category was used as training data.Uni gram approach was implemented(that is individual words in a particular tweet was compared against the training data). For each of the words in a tweet the postive, negative and neutral probability was calculated using the Naive Bayes algorithm. The three probabilities were later compared.The sentiment for the word was the highest among the three probabilities computed. The sentiment assigned to a particular tweet was positive if majority of the words in the tweet were assigned positive sentiment, negative if majority of the words in the tweet were assigned negative sentiment and

ISSN: 2278 – 1323

neutral if majority of the words in the tweet were assigned neutral sentiment. All the tweets retrieved were classified as to belonging to one of positive, negative and neutral catogory using the above approach. The final results were summarized and depicted in the form of a graph which enabled users determine the popularity of the subject queried for.

Analyzed tweets were later included into the training data, this ensured continuous learning and helped in improving the accuracy.

VI. RESULTS AND ANALYSIS

This section talks about the performance and experimental analysis of the algorithm implemented.

A. Evaluation Metric

The algorithm implemented was evaluated on the following metrics

- Accuracy
- Precision
- Recall
- F measure

Accuracy: Accuracy or Accuracy rate (or percent correct), is defined as the number of correct cases divided by the total number of cases.

Precision: Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant or it is the percentage of selected items that are correct

Recall: Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved or it is the percentage of correct items that are selected.

F Measure: A metric that combines precision and recall metrics, it is the weighted harmonic mean or can be considered as a combined measure that assesses the precision recall trade off.

Using table 1 we can calculate the above mentioned measures by using the formulas discussed below.

Accuracy = (tp + tn) / (tp + fp + fn + tn)

Precision = tp / (tp + fp)

Recall = tp / (tp + fn)

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

In the above formula P is precision, R is recall and α is factor that controls the trade off between precision and recall (0 =< α <= 1).

If we substitute for α = 0.5 (β = 1) we get the F_1 measure

$F_1 = 2 * P * R / (P + R)$

	actual class (observation)	
predicted class (expectation)	tp (true positive) Correct result	fp (false positive) Unexpected result
	fn (false negative) Missing result	tn (true negative) Correct absence of result

B. Analysis

To evaluate the Naïve Bayes algorithm implemented, the test data was extracted from the training data, 30 tweets from each of the positive and negative training data and 10 tweets from neutral training data was used as test data; hence the test data consisted of a total of 70 tweets.

The algorithm was implemented using test data as the input.

Now precision, recall and F measure were calculated as follows

- Case1: Compute precision, recall and F_1 measure for positive test data by comparing the program prediction and the actual true result.
- Case2: Compute precision, recall and F_1 measure for negative test data by comparing the program prediction and the actual true result.
- Case3: Compute precision, recall and F_1 measure for neutral test data by comparing the program prediction and the actual true result

Class	Accuracy	Precision	Recall	F_1 measure
Positive	51.43%	30%	8.33%	13.0394%
Negative	48.57%	44.44%	70.59%	54.5426%
Neutral	81.43%	14.29%	1.75%	3.1181%

Table 2 – Experimental analysis summary

ISSN: 2278 – 1323

Table 2 shows the precision, recall and F_1-measure for the test set comprising of positive, negative and neutral tweets taken from the training set.

VII. COMPARISON

The project implemented used Naïve Bayes classifier for analyzing sentiments in tweets, however sentiment analysis of tweets has been carried out by using different algorithms and different training data sets. [8] uses Support Vector Machines for classifying the twitter data, the training corpus used is obtained from Stanford twitter sentiment data. The accuracy claimed is 75.39% if the tweets are classified as to belonging to one of the 2 classes (positive or negative class) and 60.83% if the tweets are classified as to belonging to one of the 3 classes (positive or negative or neutral class). [9] uses Support Vector Machines but is focused on a specific target, the training corpus used contains tweets specific to a particular subject { Obama, Google, iPad, Lakers, Lady Gaga}. The accuracy claimed is 85.6% if the tweets are classified as to belonging to one of the 2 classes (positive or negative class) and 68.3% if the tweets are classified as to belonging to one of the 3 classes (positive or negative or neutral class).

VIII. CONCLUSION

Analysing tweets helps in determining the popularity of a subject. Analysing the tweets has various advantages for example a person who wants to know the popularity of a particular automobile can consider using this application. Based on the result of the tweet analysis obtained he/she can understand the popularity of the automobile among other fellow users. Positive summary would suggest that the automobile is well accepted in the society and the negative summary would suggest the opposite.

This paper discussed one of the possible approaches(*Naïve Bayes*) to twitter sentiment analysis. The accuracy of the project mainly depended on the quality and content of the training data.Initial accuracy was close to 50 % but adding the analyzed tweets to the training data continuously over a period of time increased the accuracy to 70%.

IX. FURTHER ENHANCEMENTS

Some of the future enhancements could be

- To help determine the popularity of a subject, data has to be gathered from various other sources such as facebook so that the accuracy of the analysis is improved.
- Bi-grams and tri-grams can be used for analysis instead of uni-grams to improve the accuracy.

X. REFERENCES

[1] Bing Liu. "Sentiment Analysis and Subjectivity". In: Handbook of Natural Language Processing , 2010, Ed. by N Indurkhya and F JEditors Damerau.

[2] R D Groot, Master Thesis on "Data Mining for Tweet Sentiment Classification" Department of Information and Computing Sciences, Utrecht University (2012).

[3] Alexander Pak, Patrick Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining" (2010). .

[4] Von Dang, Thanh Tung, "Evaluation of Natural Language Processing Techniques for Sentiment Analysis on Tweets", October 2012, pp. 3 -6.

[5] Witten, I. H., & Frank, E. Data mining: Practice Machine Learning Tools and Techniques 2nd Edition, year 2005. San Francisco: Morgan Kaufmann Publishers.

[6] Cortes, C., & Vapnik, V. "Support-Vector Networks". Machine Learning, 1995, pp. 273-297

[7] Hsu, C.-W., & Lin, C.-J. A comparison of methods for multiclass support vector machines. IEEE TRANS. NEURAL NETWORKS, 2002, pp. 415-425.

[8] Saif, H., He, Y., & Alani, H. Alleviating Data Sparsity for Twitter Sentiment Analysis. Workshop: The 2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at World Wide Web (WWW) 2012. Lyon, France, 2012.

[9] Jiang, L., Yu, M., Zhou, M., Liu, X., & Zhao, T. (2011). Target-dependent Twitter Sentiment Classification. Computational Linguistics, 151-160.

[10] Twitter4j package documentation URL: http://twitter4j.org/en/.

ISSN: 2278 – 1323

XI. BIBLIOGRAPHY OF AUTHORS

Kushwanth graduated from R.V College of Engineering in 2013 with a computer science degree. He is presently Member of technical staff at VMware R & D India. His main research interests are in the areas of Machine Learning, Natural Language Processing and Data Analytics.

Sachin Araballi graduated from R.V College of Engineering in 2013 with a computer science degree. He is presently Member of technical staff at Oracle Data Cloud India. His main research interests are in the areas of Cloud Computing, Big Data Analytics and Search Platforms.

Dr. Shambhavi B R completed her Ph.D from Visvesvaraya Technological University in the field of Natural Language Processing. She has 11 years of experience in Academics and Industry. She is presently Associate Professor in the Department of Information Science and Engineering at B.M.S College of Engineering. Her research areas are Natural Language Processing, Information Retrieval and Text Analytics. She has to her credit 5 International Journals and 8 Conference publications. She is a life member of Indian Society for Technical Education (ISTE).

Dr. Shobha G. pursued her Ph. D. from Mangalore University. She is presently Professor and Head of Department of Computer Science and Engineering Department at R.V College of Engineering. Her main research interests are in the areas of Data Warehousing, Data Mining and Data Base Management System.

ISSN: 2278 – 1323