ISSN: 2278 – 909X
International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 7, Issue 5, May 2018

BEHAVIOR ANALYSIS OF TWITTY TEXT BASED SENTIMENT USING HYBRID TF-IDF NAÏVE BAYES

MD MANAUR HUSSAIN[1] mdmanaurhussain@gmail.com

MOHAMMAD NAWAB ALI[2] (alinawab689@gmail.com)

MS AKANKSHA SEHGAL[3]Sehgal@galgotiasuniversity.ed.in

GALGOTIAS UNIVERSITY, GREATER NOIDA (UP)

Abstract

Microbloging is aparticularly prevalent broadcast medium among the Internet fraternity these days. People share their opinions and sentiments about variety of subjects like products, news, institutions, etc., every day on microbloging websites. Sentiment analysis plays a key role in prediction systems, opinion mining systems, etc. Twitter, one of the microbloging platforms allows a limit of 140 characters to its users. This restriction stimulates users to be very concise about their opinion and twitter an ocean of sentiments to analyze. Twitter also provides developer friendly streaming API for data retrieval purpose allowing the analyst to search real time tweets from various users. In this paper, we discuss the state-of-art of the works which are focused on Twitter, the online social network platform, for sentiment analysis. We take data from 'https//api.twitter.com/1.1/users/show.json' to test machine learning and hybrid approaches for sentiment analysis on Twitter. The results of both the classifier has been found to be satisfactory while the hybrid-TFIDF outperforms the Naïve Baye's Classifiers in terms of accuracy.

I. Introduction

Sentiment Analysis, also called Opinion Mining, is one of the most recent research topics within the field of Information Processing. Textual information retrieval techniques are mainly focused on processing, searching or mining factual information. Facts have an objective component; however, there are other textual elements which express subjective characteristics. These elements are mainly opinions, sentiments, appraisals, attitudes, and emotions, which are the focus of Sentiment Analysis [5]. All of them are closely related, however, they present slight differences. This fact involves the birth of many related tasks in this new research field, such as opinion mining, subjectivity analysis, emotion detection or opinion spam detection, among others. Sentiment Analysis offers many opportunities to develop new applications, especially due to the huge growth of available information in sources such as blogs and social networks. For example, recommendations of items proposed by any recommender system can be computed taking into account aspects such as positive or negative opinions about those items. Review- and opinion-aggregation websites could collect information from different sources in order to summary or compose an opinion about a candidate, product, etc., thus replacing systems which require explicitly opinions or summaries. Question answering systems represent another field where opinions play an important role. Detection of opinion-oriented questions and possible answers, and its treatment are essential to compute good answers. Detection of subjective information is really important in fields related to argumentation where objective sentences are usually more valuable. But certainly, one of the most important fields where Sentiment Analysis has a greater impact is in the industrial field. Small and big companies, as well as other organizations such as governments, desire to know what people say about their marques, products or members [6, 7, 8, 9, 10, and 11].

Sentiment Analysis is a concept that encompasses many tasks such as extraction of sentiments, sentiment classification, subjectivity classification, opinion summarization or opinion spam detection, among others. To perform any of these activities, Sentiment Analysis has to deal with many challenges. The first one is the definition of the elements involved in this area. Thus, it is necessary to define clearly concepts such as opinion, subjectivity or emotion, however, this task is not really easy. For example, in a simple way a user opinion could be considered as a positive or negative sentiment about an entity or an aspect of that entity. On the other hand, subjectivity does not imply necessarily a sentiment but it allows expressing feelings or beliefs, and specifically, our own feelings or beliefs and our emotions. These definitions have to be represented by mathematical expressions that can be computed and used as inputs for the aforementioned activities. Accordingly, Sentiment Analysis success mainly depends on the ability to extract the necessary features of those definitions from texts to perform those tasks. Thus, Natural Language Processing (NLP) techniques are essential to achieve good results depending on the task that has to be carried out. This is another of the main challenges of this research field, along with all problems related to the adaptation of typical techniques for classifying or summarizing texts in this field, as well as the creation of new techniques and algorithms specialized on opinions. Despite the complexity and difficulty of this problem, many companies and universities are developing new tools and web services which deal with several of the issues aforementioned. These services could be included, especially for research purposes, into other applications without the need of being expert in Sentiment Analysis, such as other platforms do. Following this idea and due to the growing number of new services related to Sentiment Analysis, the aim of this work is twofold. On the one hand, to present a detailed description of a set of 15 well-known free access services focused on Sentiment Analysis. These tools might have been developed by private companies or universities, but all of them allow free access to the functionalities that will be analyzed in this work. For that reason, all of them may be especially interesting for research purposes, as it is not necessary to implement services which are already working and are free. And on the other hand, this work will assess the main functionalities from these 15 services related to Sentiment Analysis and analyze the results obtained. For that purpose, three well-known data collections in the field of Sentiment Analysis will be used. This way, this work will allow the user/researcher to have enough information about the different capabilities provided by each tool, and consequently, the user/researcher can choose the most appropriate one to be included into his own platform.

Twitter

Twitter is a social medium where the users post microblogs. It is one of the most popular websites in the world, having approximately 313 million monthly active users as of June 2016, according to their home page1 . Twitter is quite well known for the strict 140 character limit on tweets (Twitter posts). This limitation of length forces users to write posts in an informal language, often tothe-point, grammatically incorrect, with slang, typos, emoticons and abbreviations. It is also common to use Twitter-specific elements like user mentions (@username) to target another Twitter user and hashtags (#tag) to symbolize topics. Usually, all content in a tweet has counted toward this limit, but now Twitter is changing which elements count. For instance, user mentions, URLs and media like videos and images, will no longer count toward the limit. This will enable the users to write longer tweets. It will be interesting to see if and how these changes the way people compose their tweets. The fact that tweets are so to-the-point, with the challenges posed by their informal language, make them an interesting field of study within natural language modeling and sentiment analysis.

Hachette

The long-running battle between Amazon and publisher Hachette has come to a close, after the companies reached agreement over online and ebook sales after months of acrimony that pitted the world's largest online retailer against

authors, agents and publishers.Since early May Amazon has been locked in a standoff with the French publishing house after Hachette refused to give Amazon pricing control over its ebooks, which would have seen most of their digital titles discounted to less than $10 a book.

II. LITERATURE SURVEY

It was during the early 1990's that the research was started in the field of sentiment analysis. The term sentiment analysis along with opinion mining was first introduced during the year 2003, during this time the work was very much limited only to subjective detection, sentiment adjectives and interpretation ofmetaphors. J.M. Weibe [16] was a research scholar who tried to present an algorithm that was able to identify subjective characters in fictional narrative text based on regularities in the text. M.A.Hearst [17] was another research scholar that had come up with intelligent text based systems to refine the information access task, while J.M. Weibe [16] was undergoing extensive examinations to try to find out if the naturally occurring narratives and regularities with the writings of the authors and come up with an algorithm that would track the point of view on the basis of these regularities. Another experimental system came into picture, it was called PHOAKS, which was abbreviated as people helping one another to know stuff by L.Terveen [2], and this would help users to find the information on the web. This system was known to be using a combined filtering approach to recognize and reuse recommendations. A browsing method using virtual reviewers for the combined exploration of movie reviews from various viewpoints. Was developed by J.Tatemura. Morinaga et al presented a framework for mining product reputation over the internet, by working in the field of marketing and customer relationship management. This approach that was defined would collect the opinions from the users automatically from the internet and text mining techniques were to obtain the reputation of the product in the market. An unsupervised method presented by P.D. Turney[1] was used to classify the reviews using a system of thumbs up and down , which would mean thumbs up for recommended and thumbs down for not recommended. It used PMI i.e. point wise mutual Information and document level classification of sentiments to get the average semantic orientation of reviews. The accuracy rate that was obtained was 74%for about 410 reviews. In some time, Turney along with Littman tried to expand their work by presenting an approach that would find the semantic orientation of a text by calculating its statistical association by using a set of positive and negative words using LSA i.e. Latent Semantic Analysis and PMI. This method was tested with 3596 words, which included a combination of 1614 positive words and 1984 negative words and had obtained an accuracy rate of 82%. Using Standard machine learning techniques a document level sentiment classification was performed by Pang et al[3]. He along with his group mates used Maximum Entropy, naïve bayes and SVM techniques to find results for unigram and bigrams and got a accuracy rate of 82.9% using three fold cross validation for unigrams. The work that they were doing also focused on better understanding of the problems and the difficulties within the sentiment classification task. A classifier was trained using reviews from the major websites by Dave et al[8]. He got a result that showed that the higher order grams can provide better results than unigrams. M.Rushdi et al [12] discovered the sentiment analysis chore by applying SVM for testing variety of domains of dataset using various weighing schemes. They used three corpora for the experimentation including a new corpus that was introduced by them and performed 10-form as well as 3-ford cross validation for each corpus. A holistic approach that would infer the semantic orientation of an opinion word that would be based on review context and would combine multiple opinions was proposed by Ding et al[18]. It took into account implicit opinions and handles implicit features that were represented by feature indicators. Study of sentiments in comparative sentences and web context based sentiments was proposed by Murthy G. and Bing Liu [19] V Suresh et al [20] presented an approach that used stop words and gaps between stop words as the feature for sentiment analysis.

III. Reserch Method

1. Date Extraction

The twitter API named as "tweety'has been used in this research for the extraction step. The major steps involved in development of the framework for live streaming of tweets begin with setting up an account on twitter.

The tweets are filtered by two ways:

- Filter by content
- Filter by location

Because to the policies of twitter the filtering is not absolutely correct and there might be a similar tweet which doesn't lie in the filtered bandwidth. The location is done using a "location 'filter available with tweety. The location filter works on the basis of longitude and latitude of the place. A TF-IDF has to be formed where the location filter works. Proposed system will consists of various modules. Each module incorporates diverse techniques to execute its definite tasks. When a specific module concludes its task, its result cum output will become input for the next module. Finally the collective determination of each and every module will be displayed.

Pre-processing the data is the process of cleaning and preparing the text for classification. Online texts contain usually lots of noise and uninformative parts such as HTML tags, scripts and advertisements. In addition, on words level, many words in the text do not have an impact on the general orientation of it. Keeping those words makes the dimensionality of the problem high and hence the classification more difficult since each word in the text is treated as one dimension. Here is the hypothesis of having the data properly pre-processed: to reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis. The whole process involves several steps: online text cleaning, white space removal, expanding abbreviation, stemming, stop words removal, negation handling and finally feature selection. All of the steps but the last are called transformations, while the last step applying some functions to select the required patterns is called filtering [14]. Features in the context of opinion mining are the words, terms or phrases that strongly express the opinion as positive or negative. This means that they have a higher impact on the orientation of the text than other words in the same text. There are several methods that are used in feature selection, where some are syntactic, based on the syntactic position of the word such as adjective, and some are univariate , based on each features relation to a specific category such as chi square (χ^2) and information gain, and some are multivariate using genetic algorithms and decision trees based on features subsets [4]. There are several ways to assess the importance of each feature by attaching a certain weight in the text. The most popular ones are: feature frequency (FF), Term Frequency Inverse Document Frequency (TF-IDF), and feature presence (FP). FF is the number of occurrences in the document. TF-IDF is given by

$$TF - IDF = FF * Log(N/DF)$$

where N indicates the number of documents, and DF is the number of documents that contains this feature [15]. FP takes the value 0 or 1 based on the feature absent or presence in the document.

Naïve bayes

The algorithm is named after famous statistician Thomas Bayes who proposed Bayesian theorem. The Naïve bayes algorithm is also based on Bayesian theorem. This theorem assumes that all the attributes are conditionally independent to each other. In this algorithm, conditional probability for each attribute with respect to certain class level is calculated. The new document is classified using sum of probabilities for each class [12]. The classifier is easy to build and useful when there is large datasets. The classification framework is briefly discussed as follows: Suppose we have D set of tuples and each tuple has attribute vector X(x1, x2, x3 , ... xn) of n dimensions. Let there are k number of classes C1, C2, C3... Ck. The classifier predicts X belongs to Ci if

$$P\left(\frac{C_i}{X}\right) = P\left(\frac{C_j}{X}\right) \quad \text{for } 1<=j<=k, j <> i.$$

$$P\left(\frac{C_i}{X}\right) = \frac{P\left(\frac{X}{C_i}\right) P(c_i)}{P(X)}.$$

Application of Naïve Bayes

1. Text Classification- The classifier is well known for its most efficient learning capability for classification of text document [13].

The steps of implementation can be listed as:

1) The data regarding student problem is collected from social media site (Twitter).

2) Then sampling of the collected data is done through training and testing of data.

3) The preprocessing of text data is done i.e. streaming, tokenization, special character removal, stop work removal is done on text data.

4) In step 4 Extraction of features is done after preprocessing of the data.

5) In this step conversion of text data to numerical data is done using TFIDF (Term Frequency Inverse Document Frequency) approach.

6) In this step data analysis is done.

7) Application of Data Mining algorithm is done.

8) In this step testing and efficacy of algorithms is done using Naïve bayes algorithm.

IV. Result and discussion

We implemented the classifier in MATLAB using Tweeter data to store the counts of words in their respective classes. Training involved preprocessing data and applying start stop handling before counting the words. Since we were using Term Frequency Inverse Document Frequency and Naive Bayes, each word is counted only once per document.

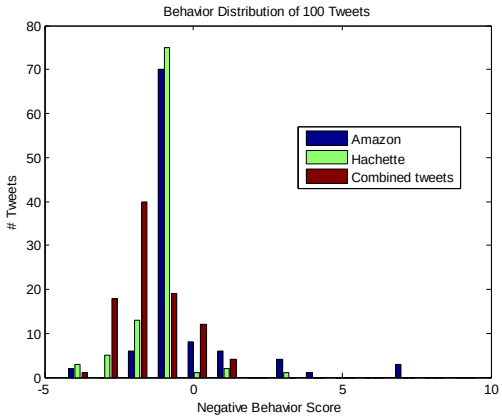

Figure 1: Behavior distribution of 100 Tweets

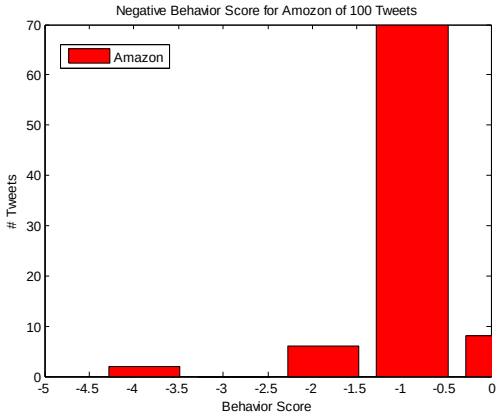

Figure 2: Negative Behavior score for Amazon distribution of 100 Tweets

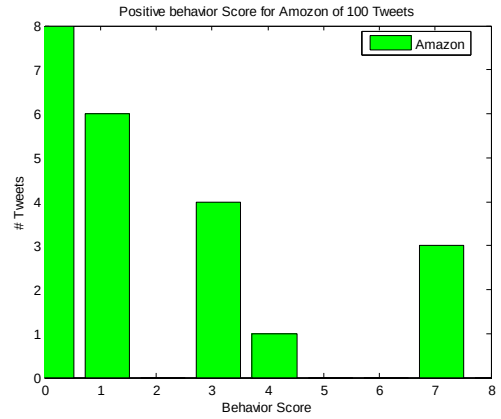

Figure 3: Positive behavior score for Amazon distribution of 100 Tweets

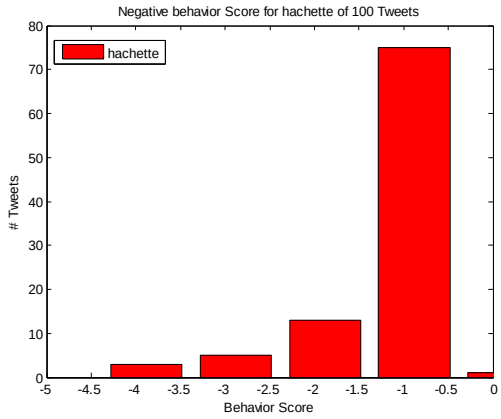

Figure 4: Negative Behavior score for Hachette of 100 Tweets

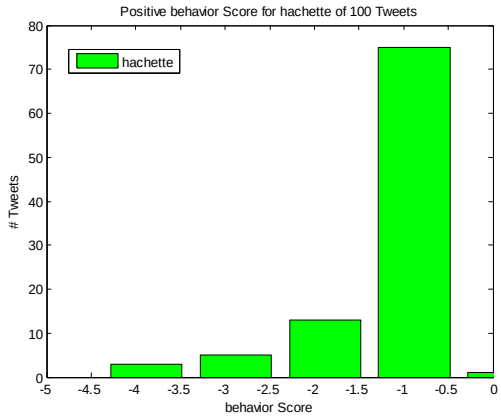

Figure 5: Positive Behavior score for Hachette of 100 Tweets

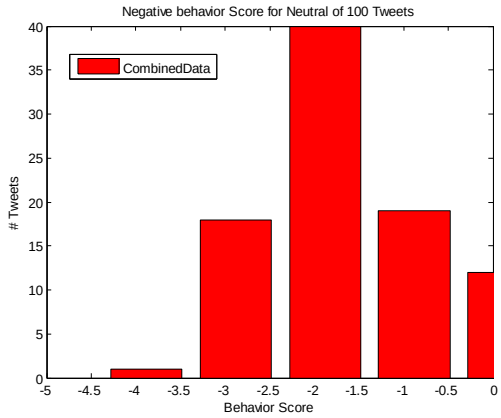

Figure 6: Negative Behavior score for Neutral of 100 Tweet

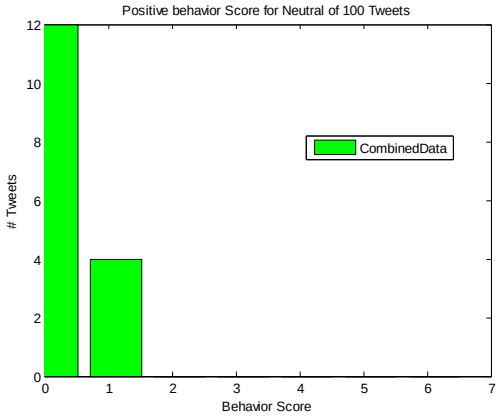

Figure 7: Positive Behavior score for Neutral of 100 Tweet

Our results show that a Hybrid TF-IDF and Naive Bayes classifier can be enhanced to match the classification accuracy of more efficient models for sentiment analysis by choosing the right type of features and removing noise by appropriate feature selection. Naive Bayes classifiers due to their conditional independence assumptions are extremely fast to train and can scale over large datasets. They are also robust to noise and less prone to over fitting.

V. Conclusion

This article has proposed a complete Twitter data analysis system to the sentiment analysis of user comments on current news items. This type of analysis has several particularities, such as its multi-focus scenario or the use of variouslanguages. Our system includes a Focus Detection Module that is able to identify the main discussion topics. It also contains a Sentiment Analysis Module which is able to analyze the strength and sentiment of the entire news article and twitty comments as well as of each of its focuses.

References

[1] P.D. Turney, "Unsupervised Learning of Semantic Orientation from a Hundred-Billion", May 16, 2002.

[2] Loren Terveen, Will Hill, Brian Amento, David Mc Donald and Josh Creter "PHOAKS: a system for sharing recommendations", March 1997

[3] B. Pang et al, "sentiment classification using machine learning techniques" 2002

[4] A. Abbasi, S. France, Z. Zhang, H. Chen, Selecting attributes for sentiment classification using feature relation networks, Knowledge and Data Engineering, IEEE Transactions on 23 (3) (2011) 447 462.

[5] B. Liu, Sentiment analysis and subjectivity, Handbook Nat. Lang. Process. 5 (1) (2010) 1–38.

[6] M. McGlohon, N. Glance, Z. Reiter, Star quality: aggregating reviews to rank products and merchants, in: Proceedings of Fourth International Conference on Weblogs and Social Media (ICWSM), 2010, pp. 114–121.

[7] A. Tumasjan, T.O. Sprenger, P.G. Sandner, M. IsabellWelpe, Predicting elections with twitter: What 140 characters reveal about political sentiment, in: Proceedings of International Conference on Weblogs and Social Media (ICWSM-2010), 2010, pp. 178–185.

[8] S.M. Mohammad, From once upon a time to happily ever after: tracking emotions in mail and books, Decis. Support Syst. 53 (4) (2012) 730–741.

[9] A. Moreo, M. Romero, J. Castro, J. Zurita, Lexicon-based comments-oriented news sentiment analyzer system, Expert Syst. Appl. 39 (10) (2012) 9166– 9180

[10] M. Castellanos, U. Dayal, M. Hsu, R. Ghosh, M. Dekhil, Y. Lu, L. Zhang, M. Schreiman, LCI: a social channel analysis platform for live customer intelligence, in: Proceedings of the 2011 International Conference on Management of Data – SIGMOD '11, ACM Press, New York, New York, USA, 2011, pp. 1049–1058.

[11] B. Chen, L. Zhu, D. Kifer, D. Lee, What is an opinion about?Exploring political standpoints using opinion scoring model, in: Proceedings of AAAI Conference on Artificial Intelligence (AAAI-2010), 2010, pp. 1007–1012.

[4] M. RushdiSaleh et al, " SVM to classify opinions in different domains", 2011

[13] Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification, Association for Computational Linguistics, 2005

[14] I. Feinerer, K. Hornik, D. Meyer, Text mining infrastructure in r, Journal of Statistical Software 25 (5) (2008) 1 54.

[15] J.-C. Na, H. Sui, C. Khoo, S. Chan, Y. Zhou, Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews, in: Conference of the International Society for Knowledge Organization (ISKO), 2004, pp. 49 54.

[16] WiebeJanyce " Identifying subjective characters in narrative, Proceedings of the International Conference on Computational Linguistics (COLING-1990)." , 1990

[17] Hearst M., 1992, Direction-based text interpretation as an information access refinement in TextBased Intelligent Systems, P. Jacobs, Editor 1992, Lawrence Erlbaum Associates, 257-274.

[18]Xiaowen Ding et al, 2008, A holistic lexicon-based approach to opinion mining, WSDM"08, February 11-12, 2008, Palo Alto, California, USA.

[19] Murthy G. and Bing Liu, 2008, Mining opinions in comparative sentences, Proceedings of the 22nd international conference on computational linguistics (Coling 2008), Manchester, August 2008, 241248.

[20] V. Suresh et al, 2011, A Non-syntactic Approach for Text Sentiment Classification with Stopwords, WWW 2011,March 28–April 1, 2011,Hyderabad, India