Aspect Based Sentiment Analysis Framework using Data from Social Media Network

Saad Ahmed
Department of Computer Science & Software Engineering,
NED University of Engineering & Technology

Saman Hina
Department of Computer Science & Software Engineering,
NED University of Engineering & Technology

Eric Atwell
School of Computing,
University of Leeds,
United Kingdom

Farrukh Ahmed
Department of Computer Science & Software Engineering,
NED University of Engineering & Technology

Abstract
Social media sites are the major source of user generated information on politics, products, ideas and services. Recently social media has become a value able resource for mining sentiment and opinions of public if the data is extracted from it reliably. In this study, a new framework is presented that uses social media network (twitter) stream data as an input and provide output in the form of identified sentiments. The main contribution of this research is a framework that employs data mining and machine learning techniques and analyzes the sentiments by using social network data. Research work has been done on social network website twitter. TF-IDF technique along with Naïve Bayes performed better (Accuracy 81.24%) in comparison with the other well-known classifiers.

Index Terms
Social networks, Sentiment analysis, TF-IDF, Data mining, Recommender system.

1. Introduction

For the purpose of data mining, different papers have been reviewed and the further reviewing of latest research papers are in the process.

In recent years, some methods of sentiment analysis have been developed in many domains; however, there is a lack of approaches that analyze the positive or negative orientation of each aspect contained in a document (a review, a piece of news, and a tweet, among others). Based on this understanding, we propose an aspect-level sentiment analysis method based on the proposed framework. The sentiment of the aspects is calculated by considering the words around the aspect which are obtained through N-gram methods. To evaluate the effectiveness of our method, we obtained a corpus from Twitter, which has been manually labeled at aspect level as positive and negative. Then it was further classified into moods and labeled as vital, alert, sure, surprise, sad and happy. The experimental results show that the best result was obtained using N-gram feature around method [1,2] with a precision of 81.93%, a recall of 81.13%, and an F-measure of 81.24%.

Since early 2000, sentiment analysis has grown to be one of the most active research areas in natural language processing (NLP) [3, 4]. It is a complicated problem with many challenging and interrelated sub-problems, including sentence-level sentiment classification. Many researchers realized that different type of sentence need different treatment for sentiment analysis. Models of different sentence types, including subjective sentences, target-dependent sentences, comparative sentences, negation sentences, conditional sentences, sarcastic sentences, have been proposed for sentiment analysis.

Certain types of documents, such as customer feedback or reviews, may contain fine-grained sentiment about different aspects (e.g. a product or service) that are mentioned in the document. This information can be highly valuable for understanding customers' opinion about a particular service or product. This is where Aspect-Based Sentiment Analysis (ABSA) comes in. With ABSA, you can dive deeper and analyze the sentiment in a piece of text toward industry-specific aspects.

The entire idea behind Aspect-Based Sentiment Analysis is to provide a way to extract specific aspects from a section of text and determine the sentiment towards each aspect separately.

Twitter, a micro-blogging website, has experienced tremendous growth in the last few years and users often post tweets related to events in real time. Users of social media tend to tweet using highly unstructured language with many typographical errors. A significant amount of tools and infrastructure are required to work on social media data due to its rapid growth and to the difficulty of processing its data by using standard relational SQL databases [5,6]. The typical change in the volume of data in social media related to a specific topic is an indication of occurrences of a real world event. Events in Twitter can be detected by clustering similar tweets. The text in social media has been often noisy, and high volume of such data renders the challenge of finding a suitable

Manuscript received July 5, 2017
Manuscript revised July 20, 2017

technique for event detection to the researchers. Today, some solutions such as Topic Detection and Tracking over Newswire system and event detection in social media have been employed in data mining related research, there remains a great potential in research for extracting sentiments from tweet clusters and trending them.Social networks plays a vital role in the mood of the public, as the information can be twitted or posted as Social networking, blogging, RSS feeds, etc. Applications are available Web 2.0. The internet users share their views, and ideas using these social networking sites. So the information is increasing every day. Social networks like Twitter or Facebook sites are most popular. Facebook reaches its 1 billion users [7] in October 2012, while twitter had more than 500 million users on February 2012 [8,9] and currently it has more than 690 million registered users [10]and on average twitter recorded 9100tweets/second.

With over 41 million tweets[11], the most discussed topic ever on Twitter was the "#ALDubEBTamangPanahon" of AlDub on October 24, 2015 during the sold out special concert presentation of the Kalyeserye segment of the noontime show Eat Bulaga entitled Eat Bulaga: Sa TamangPanahon held at the Philippine Arena, the world's largest indoor arena located in Bulacan, Philippines. The said event was attended by over 55,000 fans.

With over 35.6 million tweets, the most discussed sports game ever on Twitter was the 2014 FIFA World Cup semi-finalbetween Brazil and Germany on July 8, 2014.

The most tweeted moment in the history of Twitter was during the airing of Castle in the Sky on August 2, 2013, when fans tweeted the word "balse" at the exact time that it played in the movie. There was a global peak of 143,199 tweets in one second, beating the previous record of 33,388.

This makes twitter very popular social media site for data miners and researchers as there is huge amount of data available all the time with all kind of information [12]. The information such as followers, followed, tweets, and posts can be used in Recommender system[13,14] to extract valuable information like public interests, trends, sentiments etc [15,16,17]. ABSA emerges as excellent technique which enables us to find the best solution. Recently ABSA based on social network data is gaining importance in the field of data mining. The result of this proposed frameworkwill reflect the mood of the general public.

Methodology

The main focus of this work will be to develop a new system that employs data mining and machine learning techniques by using social network data to perform ABSA.

Figure 1 illustrates our proposed framework.

This Research comprises of following parts.

1. Web mining using twitter API. (Data Crawler)
2. Converting unstructured data stream to structured dataset and store it in a database.
3. Natural Language (NLP) Processing unit.
4. Using learning classifier algorithm
5. Find suggestions(polarity) using soft computingtechnique (K-means).

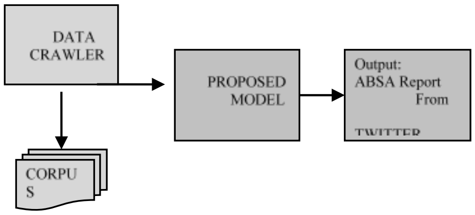

Figure 1: Proposed Framework for ABSA

Twitter, a micro-blogging online social media network is considered as a platform in this research for ABSA. Twitter generates vast amounts of data, which is exceptionally valuable for mining events and trending them. The ease of publishing messages in Twitter makes it a popular data source for real-time sentiment analysis. There are many challenges in using data for ABSA from Twitter, such as:

1. A system is needed to process the massive volume of tweets that are submitted at a fast rate.
2. A powerful keyword mining tool is required to represent the events, as tweets are very short due to their 140 character length limitation.
3. A well-organized technique is needed to handle unstructured, noisy and grammatically incorrect text messages.

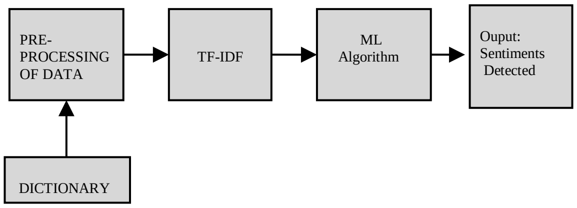

Figure 2: Proposed Model to perform ABSA.

Figure 2 illustrates the proposed framework using TF-IDF technique [18]. The Twitter data corpus is divided into chunks of equal size to achieve a constant processing time for the testing process. The process is detailed in the following sub sections.

Data Preprocessing

The data was collected from social media website (Twitter) by using the API (application Program Interface) and the Data Crawler. The downloaded dataset has 40,000+ tweets(document). Then this data was pre-processed, in which the data cleaning is done, removal of garbage such as links of image and website addresses which are of no use in this research project.

The downloaded dataset has very detailed information about each tweet. Only the relevant and necessary fields of data was collected and stored in .csv files, the fields includes User ID, User Screen Name, Reference, Tweet ID, Date and Time remaining data was filtered out. Now this pre-processing step is completed. Dataset is in structured form and ready to be tested.

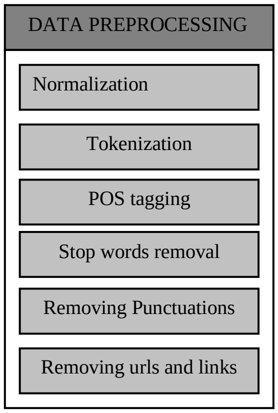

Figure 3: Main steps of data/text preparation

Text Preparation

The tweets were parsed into a corpus for text analysis. The following steps were performed to clean the corpus and prepare it for further analysis. Only the text portion of the tweet is considered. Figure 3 illustrate the main steps of text/data preparation

Removing numbers: Tweet IDs are number generated by Twitter to identify each tweet. Numbers as such des not serve any purpose for text analysis and hence they are discarded.

Removing URLs & links: Many tweets contained links to webpages and videos on the Internet. These were also removed.

Removing stopwords: Stopwords are words in English that are commonly used in every sentence, but have no analytical significance. Examples are ‘is’, ‘but’, ‘shall’, ‘by’ etc. These words were removed by matching the corpus with the stopwords list in the tm package of R. Expletives were also removed.

Stemming words: In text analysis, stemming is ‘the process of reducing inflected (or sometimes derived) words to their word stem, base or root form—generally a written word form’. Stemming is done to reduce inflectional forms and sometimes derivationally related-forms of a word to a common base form. Many methods exist to stem words in a corpus.

Suffix-dropping algorithms: The last parts of all the words get truncated. For example, words like ‘programming’, ‘programmer’, ’programmed’, ‘programmable’ can all be stemmed to the root ‘program’. On the other hand, ‘removing’, ‘remove’, ‘removed’ are stemmed to form ‘remov’, which is not a word or a root. This method was used for simplicity.

Lemmatisation algorithms: Each word is the determination of the *lemma* for a word in the corpus. This is done with the understanding of the context, part of speech and the lexicon for the language. For example, ‘better’ is related to ‘good’, ‘running’ is related to ‘walk’ and so on.

Removing punctuation: Punctuation marks make no impact to the analysis of text and are hence removed.

Stripping whitespace: Words that have extra whitespaces at the beginning, middle or end are subjected to a regular expression that removes the whitespace and retains only the words themselves.

During preprocessing of the tweets a considerable level of noise is removed. This is done by using tokenizing. Tokenizing is a process of splitting text into a set of individual terms or tokens. Each tweet is tokenized into a sequence of terms. Only tweets written in the English language are considered for clustering. In Natural Language Processing the most commonly used words in a text document are referred as stopwords. The qualified tweet is checked against a standard stopword list. The Natural

Language Tool Kit's stopword list is applied in this process to eliminate terms which carry the least information. For example, articles like 'the', 'a', and 'an' are removed from the tweets. Also URLs and the token starting with '@' (i.e. a mention or reply) will be removed from the tweets in the filtration process. At the end of this process, each tweet is split into a set of features which are present in the vector space model.

Term Frequency-Inverse Document FrequencyTechnique

TF-IDF is a solution to the unsupervised document clustering problem. In TF-IDF based approaches the document set to be clustered must be known in advance or otherwise, high computational complexity is required. Most of these existing methods work under the assumption that the whole data set is available and static. For instance, in order to use the popular Term Frequency – Inverse Document Frequency approach and its variants, one needs to know the number of documents in which a term occurred at least once (document frequency). This requires a priori knowledge of the data, and that the data set does not change during the calculation of term weights.

The need for knowledge of the entire data set significantly limits the use of these schemes in applications where continuous data streams must be analyzed in real-time. For each new document, this limitation leads to the update of the document frequency of many terms and therefore, all previously generated term weights needs recalibration

In Term Frequency – Inverse Corpus Frequency it does not require term frequency information from other documents within the set and thus, it can process document streams in linear time.

In this proposed framework, we have used TF-IDF technique which provides us term weights. This technique is used to filter out tweets which have minimal or no information to predict the sentiments. In numerical analysis, a sparse matrix is a matrix in which most elements are zero; this implies that they have many parameters which are not very informative. Therefore, to reduce sparsity, the terms that occur frequently were removed. This tends to have the effect of both reducing over fitting and improving the predictive abilities of the presented frame work. In this case, we have limited the sparsity of the matrix to 70% in order to acquire better results without any biasness to the context of the public sentiment.

The tweets containing the important terms were retained and remaining tweets were filtered out form the dataset as these tweets have no information to predict the sentiments and by doing this we improve the performance of the ML algorithm.

Most of the datamining frameworks uses clustering methods. Clustering method groups similar items in a one subset [19]. These subsets of items are called clusters. Clustering uses different types of techniques such as Nearest Neighbour (NN) and K-mean. In the presented study, the clusters which are formed using K- Means will be used as different polarities. The cluster which has the highest number of items will be considered as the sentiment polarity of that tweet.

Experimental Results

The dataset is divided into two equal parts i.e. training and testing data. The presented framework was tested using different ML algorithms with TF-IDF, using the dataset acquired from the twitter online.Figure 4illustrate the performance of different ML algorithms. The performance of Naive Bayes algorithm in our proposed framework was exceptional. Figure 5 illustrate the output predicted by the proposed frame work.

Table 1: Performance of presented framework using different algorithms

Algorithm	Accuracy	Precision	Recall	F Score
Nave Bayes	0.811	0.816	0.815	0.824
SVM	0.721	0.665	0.690	0.670
Forest	0.812	0.765	0.765	0.765
Tree	0.314	0.150	0.500	0.230
Bagging	0.608	0.520	0.520	0.520
Maxen	0.331	0.335	0.310	0.295

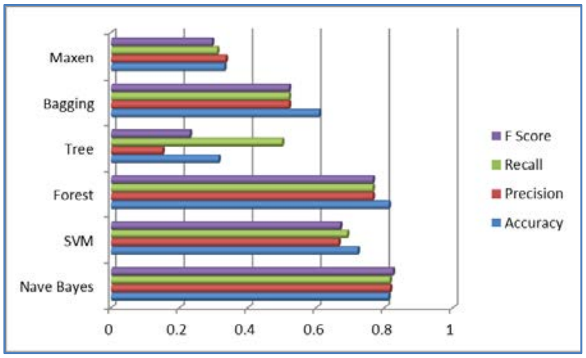

Figure 4: Performance of different ML Algorithms

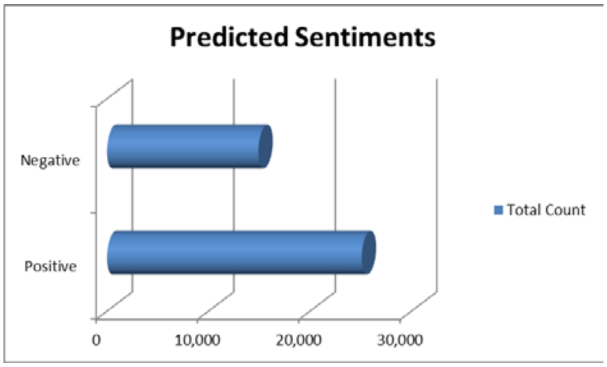

Figure 5: Sentiments predicted by the framework

Conclusion

Social media networks provide an important source of information regarding users andtheir sentiments towards a particular product, service or event. This is especially valuable in giving in depth knowledge about the current trends and moods of public. In this paper we have presented a formal framework to analyze sentiments from social media sites (twitter). This framework has been evaluated on 40k tweets and further evaluation will be doneby increasing data set to provide efficient/faster processing on big data.

The best results were obtained when Naïve Bayes classifier algorithm is used which gave 81.1% accuracy on the given dataset of tweets. Real data from twitter website was received using the Twitter streaming API. The API provides extensive metadata for the tweets.

Data pre-processing has been done to improve the accuracy of the proposed framework. The framework which we have proposed uses TF-IDF and naive bayes machine learning algorithm and is expected to detect the sentiments faster with better F-scores in comparison with the other well-known classifiers.

References

[1] Cavnar, William B., and John M. Trenkle."N-gram-based text categorization."Ann Arbor MI 48113.2 (1994): 161-175.

[2] Salas-Zárate, María del Pilar, et al. "Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach." Computational and mathematical methods in medicine 2017 (2017).

[3] Liu, Bing. Sentiment analysis: Mining opinions, sentiments, and emotions. Cambridge University Press, 2015.

[4] Ravi, Kumar, and Vadlamani Ravi. "A survey on opinion mining and sentiment analysis: tasks, approaches and applications." Knowledge-Based Systems 89 (2015): 14-46.

[5] J. Lin and D. Ryaboy, "Scaling big data mining infrastructure: the twitter experience," SIGKDD Explor.Newsl., vol. 14, pp. 6-19, apr, 2013.

[6] G. Mishne, J. Dalton, Z. Li, A. Sharma and J. Lin, "Fast data in the era of big data: Twitter's real-time related query suggestion architecture," in Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, USA, 2013.

[7] Ashlee Vance (October 04, 2012) Facebook: The Making of 1 Billion Users. [Online]. Available: http://www.businessweek.com/articles/2012-10-04/facebook-the-making-of-1-billion-users

[8] Lauren Dugan (February 21, 2012) News, Statistics: Twitter to Surpass 500 Million Registered Users on Wednesday. [Online]. Available : http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842

[9] Twitter Inc. (2011) Year in Review: Tweets per second. [Online]. Available: http://yearinreview.twitter.com/en/tps.html

[10] http://www.statisticbrain.com/twitter-statistics/, Access date: 10-03-17, 2017.

[11] "Fans in the Philippines& around the world sent 41M Tweets mentioning #ALDubEBTamangPanahon". Twitter Data Verified Account. October 27, 2015. Retrieved October 30, 2015.

[12] J. Bobadilla, F. Ortega, A. Hernando, A. Gutierrez, Recommender systems survey Universidad Politécnica de Madrid, Ctra. De Valencia, Km. 7, 28031 Madrid, Spain

[13] E.R. Nunez-Valdez, J.M. Cueva-Lovelle, O. Sanjuan-Martı´nez, V. Garcı´ a-Dı´az, P. Ordonez, C.E. Montenegro-Marı´ n, Implicit feedback techniques on recommender systems applied to electronic books, Computers in Human Behavior 28 (4) (2012) 1186–1193.

[14] J. Bobadilla, F. Ortega, A. Hernando, A. Gutierrez, Recommender systems survey Universidad Politécnica de Madrid, Ctra. De Valencia, Km. 7, 28031 Madrid, Spain

[15] K. Choi, D. Yoo, G. Kim, Y. Suh, A hybrid online-product recommendation system: combining implicit rating-based collaborative filtering and sequential pattern analysis. Electronic Commerce Research and Applications, in press, 2012.02.004.

[16] S.K. Lee, Y.H. Cho, S.H. Kim, Collaborative filtering with ordinal scale-based implicit ratings for mobile music recommendations, Information Sciences 180 (11) (2010) 2142–2155.

[17] E.R. Nunez-Valdez, J.M. Cueva-Lovelle, O. Sanjuan-Martı´nez, V. Garcı´ a-Dı´az, P. Ordonez, C.E. Montenegro-Marı´ n, Implicit feedback techniques on recommender systems applied to electronic books, Computers in Human Behavior 28 (4) (2012) 1186–1193.

[18] Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. "A comparative study of TF* IDF, LSI and multi-words for text classification."Expert Systems with Applications 38.3 (2011): 2758-2765.

[19] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.