International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 12, December-2015

A Survey on Classification of Sentiments from Twitter

Atul Patil
MIT College of Engineering,
Kothrud, Pune, India

Kunal Kulkarni
MIT College of Engineering,
Kothrud, Pune, India

Vinit Magar
MIT College of Engineering,
Kothrud, Pune, India

Ajay Manwar
MIT College of Engineering,
Kothrud, Pune, India

***Abstract*— Sentiment Analysis or Opinion Mining is the computational study of the attitudes, opinions and emotions of people toward an entity. The 21st century has seen the advent of the online shopping system, which has become a very popular mechanism for users to purchase goods without actually going to stores. These reviews can be referred to by other users to help them decide which products to buy. Applying sentiment analysis to these product reviews can help in understanding the overall reception of the product. Sentiment analysis has increased the interest of many researchers in recent years, since subjective texts are useful for many applications. Sentiment analysis in particular has become a great research field. Studies on sentiment analysis mainly focus on pre-processing, feature extraction and classification model construction. This paper presents a survey on various research and latest development done in sentiment analysis, and also makes an introduction of its application in various fields. And finally discuss some future scope and directions where sentiment analysis could grow.**

Keywords*— Nonlinear system; generalized predictive control; support vector machines

I. INTRODUCTION

Sentiment analysis, also known as opinion mining is a type of natural language processing for finding out the overall attitude of the public about a particular product, movie or any such entity. Sentiment analysis involves the collection and examination of opinions about the product or the movie made by people in various online media like blogs and tweets. Sentiment Analysis can be highly useful for a wide range of people. For example, Sentiment Analysis done in the field of marketing can help a person judge the success of an ad campaign or decide which versions of a product or service are popular and also find out which demographics like or dislike certain features of the product.

There are several challenges in Sentiment analysis. People do not always express their opinion clearly on the Internet. A lot of comments are replete with the use of slang words. Not only that, but users can post their opinions in a wide variety of languages, depending upon their preference. Moreover, an opinion word which can be considered to be positive in a certain context may be considered negative in another. For example, "the product is great" is very different from "the product is not great". Sometimes, reviews may also contain both positive and negative sentiments. This can be managed by analyzing one sentence at a time in the case of large documents. However, in case of tweets and blogs, which have a certain limit on the maximum allowed characters, it becomes difficult, since the text lacks context.

Three key areas of research in Sentiment Analysis are sentiment classification, feature based Sentiment classification and opinion summarization. In sentiment classification, entire documents are classified based on the opinions towards a certain object. Feature based classification is done by considering certain features of the objects. In opinion summarization, only the features of the product are mined on which the customers have expressed their opinions.

Languages that have been mostly studied are English and Chinese. Presently, there are very few researches conducted on sentiment classification for other languages like Hindi, Arabic and Italian. This survey aims at focusing towards the work in English. The emergence of sentiment analysis dates back to late 1990's, but it has become a major emerging sub field of information management from 2000, which this survey focuses.

II. DATA SOURCES

Since user's opinions must be analyzed for the sentiment analysis process, data sources where a large number of users are active must be analyzed. Blogs, Twitter and review sites can provide a good understanding of the overall sentiments of a particular product.

A. Blogs

With the advent of the Internet, blogging has grown rapidly. Blog pages are one of the most popular media used by people to express their opinions. A lot of these blogs contain reviews on many products, movies and issues. Many studies related to sentiment analysis use blogs as the source of opinion (Martin, 2005; Murphy, 2006; Tang et al., 2009).

B. Twitter

Twitter is a social media in which users can post tweets limiting to 140 characters. Users can express their opinion regarding a wide variety of issues or products. This has become a rich source of data for sentiment analysis in recent times.

C. Review Sites

Opinions of others can be important factors in helping a customer decide whether to buy a product or not. In a majority of the studies, the data is collected from websites like www.amazon.com for product reviews, www.yelp.com for restaurant reviews and www.reviewcentre.com, again for product reviews.

III. RELATED WORK

Sentiment Analysis has been described as a Natural Language Processing task at many levels of granularity. Starting out from being mapped into a document level classification task[10], then applied sentence level and phrase level classification.

A. Data Gathering

Extraction of tweets is done by retrieving data related to the target, by going through the whole dataset and extract all the tweets which contain the keywords of the target[8]. Twitter's API provides a straightforward way to query for users and returns results in a JSON format which makes it easy to parse. While gathering data an error message "rate limit exceeded" is often encountered. This is because Twitter imposes a limit on the number of API calls a single app can make in set "window" of times (15 minutes). For solution to this problem there are two approaches, by either making multiple Twitter Apps and request additional OAuth credentials or set up a cronjob (scheduled) task to run every 15 minutes. Doing so will enable script to run during scheduled times or intervals in the background.

B. Machine Learning

Overall, text classification using machine learning is a well-studied field.[10] Various machine learning techniques (Naive Bayes, Maximum Entropy, and Support Vector Machines) in a specific domain. Accuracy of each technique is observed in these studies [3]. It describes support vector machine (SVM) as a concept in statistics and computer science for a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. Furthermore, it makes use of Least Squares support vector machine where the inequality constraints of standard SVM are replaced by the equality constraints and solve quadratic programming problems with the way of solving linear equations.

Confirmed studies are held to get higher accuracy rate for sentiment analysis [7] . These studies were able to implement such analysis using various types of machine learning algorithm and feature extractors. The best candidates for sentiment analysis are SVM (Support Vector Machine), Naïve Bayes and Maximum Entropy. According to (Turney, 2002), accuracy is strongly influenced by the context in which the words are used.

It has also been stated that unsupervised learning generally functions as follows: in the initial training phase, an inductive process learns the characteristics of a class based on a feature set of pre classified documents (reference corpus) and it then applies the acquired knowledge to categorize unseen documents, during testing.

C. Text Sentiment

Twitter becomes almost an indefinite source in text classification. In a large scale study, over 1,600,000 Tweets have been downloaded and classified into positive and negative classes\cite{Online Jobs}.Twitter users post messages from many different media, including their cell phones. The frequency of misspellings and slang in tweets is much higher than in other domains. Various attributes are considered while processing tweets.

- Length - The maximum length of a Twitter message is 140 characters
- Data availability - Another difference is the magnitude of data available. With the Twitter API, it is very easy to collect millions of tweets for training. In past research, tests only consisted of thousands of training items.
- Domain - Twitter users post short messages about a variety of topics unlike other sites which are tailored to a specific topic. This differs from a large percentage of past research, which focused on specific domains such as movie reviews.

D. Stop words

(El-Khair, 2006) stated that stopwords are very common words that appear in the text that carry little meaning. They serve only a syntactic function but do not indicate subject matter. The removal of the stopwords also changes the document length and subsequently affects the weighting process.The removal of the stopwords can increase the efficiency of the indexing process as 30 to 50% of the tokens in a large text collection can represent stopwords.

E. Support Vector Machine

Support vector machines have been shown to be highly effective at traditional text categorization, generally outperforming Naive Bayes \cite{pang}. They are large-margin, rather than probabilistic, classifiers, in contrast to Naive Bayes and MaxEntropy. Support Vector machine is a kind of vector space model based classifier which requires that the text documents should be transformed to feature vectors before they are used for classification. Possible factors affecting the SVMs accuracy are the choice of kernel, the extraction of features and the weighing of features. Usually the text documents are transformed to multidimensional tf.idf vectors. Every text document is classified represented as a vector into a particular class.

F. Optimization and Pre-processing

The use of Support Vector Machine can be optimized for classifying positive or negative feedback [1]. This can be done by through choice of kernel and proper tuning of SVM hyper-parameters as core factors in contributing to SVM accuracy, having huge amount of training data to widen hyper plane of vectors. The Sentiment classification is done by first gathering the data using Twitter API. The Tweets are gathered via subjects (the brand name), they can be any kind of tweets re-tweets, mentions of other users, links, smileys.

The next stage is pre-processing stage where the sentiments that are gathered are filtered. It involves conversion of characters to lower case, detecting and removing any URLs or links in the tweet, removing username referred by @ symbol, punctuation and white space as they don't make any significant changes to the meaning of the tweet.
It also involves conversion of data into SVM format
<label><index0>:<tf-idf><index1><tf-idf>...<indexn><tf-idf>

IV. APPLICATIONS OF SENTIMENT ANALYSIS

Sentiment Analysis finds its applications in a wide variety of fields like marketing of products, predicting the results of elections beforehand, by analyzing public sentiment. Online advertising has been increasing in recent times and Sentiment Analysis is largely used in this field. Guang Qiu(2010) focuses on Dissatisfaction oriented online advertising, whereas (Teng-Kai Fan, Chia-Hui Chang ,2011) focuses on Blogger-Centric Contextual Advertising.

For the purpose of identifying potential risks, companies need to collect and analyze information about the products and plans of their competitors. Competitive intelligence (Xu, 2011) makes use of sentiment analysis to extract and visualize comparisons and relations between products from customer reviews, with the inter dependencies among relations taken into consideration. This helps enterprises discover potential risks and then design new products and marketing strategies.

Other applications include online message sentiment filtering, mail sentiment classification, attitude analysis of authors of web blogs etc.

V. EVALUATION OF PERFORMANCE

Most Sentiment Analysis algorithms would categorize the data into Positive / Neutral / Negative. So, the rule of thumb is to measure performance is whether the system categorized the data in accordance with the intuition of the user. This a very abstract as well as subjective problem, whose accuracy cannot be measured by plain mathematics. If yes, the system is accurate, else it is not. Essentially, this is more of an agreement rate, which can be expressed by

A = (No. of Expected Outcomes) / (Total No. of Runs).

To evaluate the performance of sentiment classification, this paper has adopted precision, recall and F-Measure as a performance measure.

Recall (also known as sensitivity) is the fraction of relevant instances that are retrieved. It is the measure of completeness, or sensitivity, of a classifier. Higher recall indicates less false negatives, while lower recall indicates more false negatives. Recall improvement can often decrease precision because it gets increasingly harder to be precise as the sample space increases.

$$\text{Recall} = \frac{number\ of\ correct\ positive\ predictions}{number\ of\ positive\ examples}$$

Precision (also called positive predictive value) is the fraction of retrieved instances that are relevant. It is the measure of the exactness of a classifier. Higher precision indicates less false positives, while a lower precision indicates more false positives. It is often at odds with recall, as an easy way to improve precision is to decrease recall.

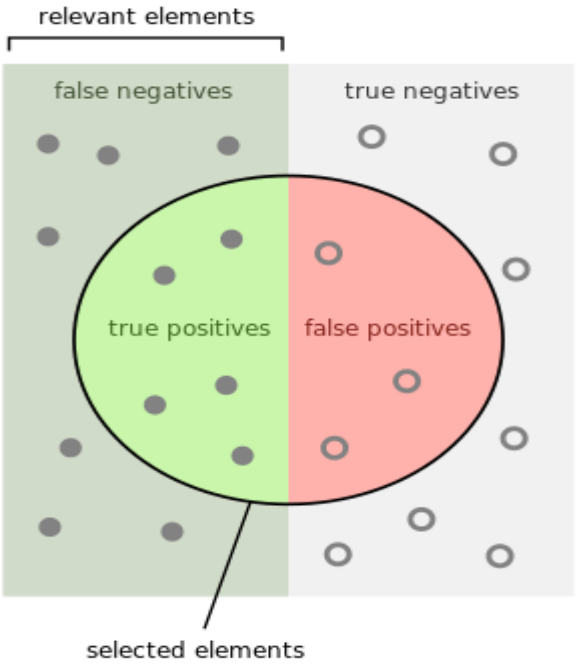
$$\text{Precision} = \frac{number\ of\ correct\ positive\ predictions}{number\ of\ positive\ predictions}$$

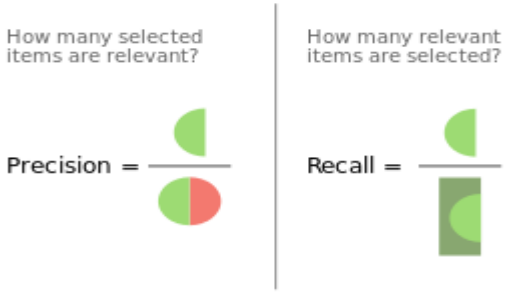
F-measure is a single metric which is a combination of precision and recall, which is the weighted harmonic mean of precision and recall.

$$\text{F-measure} = \frac{2 * Recall * Precision}{Recall + Precision}$$

G-measure also is a measure evaluated with the combination of precision and recall, except unlike F-measure, it does not considers harmonic mean, instead it considers the geometric mean.

$$\text{G-measure} = \sqrt{precision * recall}$$





VI. CONCLUSION AND FUTURE SCOPE

With the rise of the Internet and social networking giants like Facebook and Twitter, interest in the field of sentiment analysis has grown tremendously over the past few years. Sentiment analysis has numerous applications, some of which have been discussed above. The study of various approaches to sentiment classification reveals that neither classification model consistently outperforms others. Different models can be combined together to maximize the benefits of all such models.

In the future, a lot of work can be done by considering other regional languages like Hindi, Marathi, French, Spanish etc. Moreover, the limitations of the current approaches can be identified and research can focus on these limitations, thus enhancing the whole process of sentiment classification. A lot of users try to include sarcasm in their posts, which may not be accurately classified with the current models. The use of negation expressions may give rise to difficulties in sentiment classification. Thus, more future research could focus on these areas and help classify sentiments accurately and efficiently.

REFERENCES

[1] Allen Banados, Jao, and Kurt Junshean Espinosa. "Optimizing Support Vector Machine in classifying sentiments on product brands from Twitter." *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on. IEEE, 2014.*

[2] Miranda, Marcelo Drudi, and Renato José Sassi. "Using Sentiment Analysis to Assess Customer Satisfaction in an Online Job Search Company." *Business Information Systems Workshops. Springer International Publishing, 2014*

[3] Xu, Yong. "Generalized predictive control model based on support vector machines." *Natural Computation (ICNC), 2012 Eighth International Conference on. IEEE, 2012.*

[4] Li, Lei, Zhi-ping GAo, and Wen-yan Din. "Fuzzy multi-class support vector machine based on binary tree in network intrusion detection." Electrical and Control Engineering (ICECE) , *2010 International Conference on. IEEE, 2010.*

[5] Bifet, Albert, and Eibe Frank. "Sentiment knowledge discovery in twitter streaming data." Discovery Science. Springer Berlin Heidelberg, 2010

[6] Go, Alec, RichaBhayani, and Lei Huang. "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford 1 (2009): 12.*

[7] Gamon, Michael. "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis." *Proceedings of the 20th international conference on Computational Linguistics. Association for Computational Linguistics, 2004*

[8] Tan, Shulong, et al. "Interpreting the public sentiment variations on twitter." *Knowledge and Data Engineering, IEEE Transactions on 26.5 (2014): 1158-1170*

[9] Nirmal, V. Jude, and DI George Amalarethinam. "Parallel Implementation of Big Data Pre-Processing Algorithms for Sentiment Analysis of Social Networking Data."

[10] Pang, Bo; Lee, Lillian; Vaithyanathan, Shivakumar (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques" *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 79–86.*