International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 08 | Aug-2016 www.irjet.net p-ISSN: 2395-0072

Sentiment Analysis Using SVM and Maximum Entropy

Snehal L. Rathod[1], Sachin N.Deshmukh[2]

[1]M.Tech, Dept. of CS and IT, Dr. BAMU, Maharashtra, India
[2]Assistant Professor, Dept. of CS and IT, Dr. BAMU, Maharashtra, India

---

Abstract - *With the rapid proliferation of online-blogging and micro-blogging web sites, hundreds of thousands of text posts are generated and made to be had online daily. Utilizing this affluent abstracts approach could facilitate accomplished purchasing of items, advertent trends and accessible tendencies apropos assorted articles accessible in the market, advertent political affection of societies above-mentioned to a national election, etc. Using this rich information could facilitate educated purchasing of objects, discovering developments and public developments involving more than a few merchandise in the market, discovering political inclination of societies previous to a country wide election, and many others. Considering that the final decade, Sentiment evaluation (SA) has received increased attention from many researchers as a procedure for addressing subject matters, such as the a forementioned ones. This paper specializes in SA making use of sentiment Features. We proposed exclusive sentiment polarity detection approaches. In our experiments, we appearance that our polarity apprehension methods are highly effective and can beat the above baselines in a lot of our conducted experiments.*

Key Words: **Sentiment analysis, support vector machine, maximum entropy, artificial intellengence, with features, without features, artificial intelligence**

1. INTRODUCTION

In recent years, we now have witnessed that opinionated postings in social media (e.g. stories, discussion board discussions, blogs, micro-blogs, Twitter, feedback, and postings in social network websites) have helped reshape corporations, and sway public sentiments and emotions, which have profoundly impacted on our social and political methods [2]. There is a area of be trained that analyzes this kind of opinionated postings which most of the time referred to as as sentiment evaluation or opinion mining [2]. The dataset used in this research is a group of tweets. Those tweets shall be extracted and processed so it will probably produce information such as sentiment that consists in tweets. Sentiment analysis on tweets is used to find out whether a tweet consists of positive or negative sentiment. There are two kinds of learning that usually used in the process of sentiment analysis, which is supervised learning and unsupervised learning [3]. The machine learning system is belong to supervised studying, this procedure quite often want so much if training information that have been labeled manually. Without labeled training data, supervised learning won't capable to be processed [3][4]. Whereas, the lexicon-centered method is belong to unsupervised finding out, which does not want already trained information and most effective depend on the dictionary that's used [3].

As mentioned above ways have special characteristics, however it may complementary if both methods are mixed. The combination of each methods can be finished by way of utilising lexicon-based system to create labeled tweets which can be utilized as training information in Support Vector Machine process so there shall be no training approach on this blend ways[5][3]. The carry of online social media in the latest years has converted the best way for peoples be in contact with each other when they share ideas and opinions.

Micro blogging web sites such as Twitter have gained increased repute, and as a consequence, novel and yet wealthy data channels are formed. Every day a big amount of formal or informal texts are made available on line through this online media. The knowledge captured from these texts, could be active for scientific surveys from a amusing or political perspective [6]. On one hand, companies and product owners who aim to alleviate their products/services may strongly account from the affluent acknowledgment [7], [8]. On the opposite hand, customers would also gain knowledge of about positivity or negativity of customers respecting distinctive elements of merchandise/services and therefore may just react.

Sentiment analysis (SA) is the system of extracting the polarity of individuals's subjective opinions from plain normal language texts [9].

An SA procedure, takes as enter a collection of files with unknown polarity, and returns as output the opinions expressed in those records and their envisioned polarity. This makes it possible for both consumers and organizations to have convenient access to public's opinion concerning distinct gadgets/merchandise. There was a fine amount of earlier study on quite a lot of methods using the net technology to maximize the benefits of customers, as good as, firms available in the market location [10].

Twitter messages as many others posted on the blogosphere are in most cases informal. On account that of the anomalistic nature of informal textual content, evaluation or processing of this sort of text is more often than not more difficult when compared to formal textual content. The important change between processing formal and informal text is in knowledge preprocessing. Formal text probably wishes less preprocessing. Casual textual content however, more often than contains emoticons, use of slangs, not good grammar, and sarcasm or non-dictionary-usual words. Therefore, analysis of this category of textual content is usually more complicated. In this paper, we endorse quite a lot of ways for handling both formal and informal texts, and compare the offered methods against two baseline approaches in a benchmark.

Sentiment Analysis: Opinion mining is a type of natural language processing for tracking the mood of the public about a particular product. Opinion mining, which is also called sentiment analysis, involves building a system to collect and categorize opinions about a product. Automated opinion mining often uses machine learning, a type of artificial intelligence (AI), to mine text for sentiment.

Opinion mining can be useful in several ways. It can help marketers evaluate the success of an campaign or new product launch, determine which versions of a product or service are popular and identify which demographics like or dislike particular product features. For example, a review on a website might be broadly positive about a digital camera, but be specifically negative about how heavy it is. Being able to identify this kind of information in a systematic way gives the vendor a much clearer picture of public opinion than surveys or focus groups do, because the data is created by the customer.

There are several challenges in opinion mining. The first is that a word that is considered to be positive in one situation may be considered negative in another situation. Take the word "long" for instance. If a customer said a laptop's battery life was long, that would be a positive opinion. If the customer said that the laptop's start-up time was long, however, that would be is a negative opinion. These differences mean that an opinion system trained to gather opinions on one type of product or product feature may not perform very well on another.

A second challenge is that people don't always express opinions the same way. Most traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much. In opinion mining, however, "the movie was great" is very different from "the movie was not great".

Finally, people can be contradictory in their statements. Most reviews will have both positive and negative comments, which is somewhat manageable by analyzing sentences one at a time. However, the more informal the medium (twitter tweets or blog posts for example), the more likely people are to combine different opinions in the same sentence. For example: "the movie bombed even though the lead actor rocked it" is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone though based on a short piece of text because it lacks context. For example, "That movie was as good as his last one" is entirely dependent on what the person expressing the opinion thought of the previous film.

Different Levels of Analysis: In general, sentiment analysis has been investigated mainly at three levels:

Document level: The task at this level is to classify whether a whole opinion document expresses a positive or negative sentiment. This stage of analysis assumes that every report expresses opinions on a single entity (e.g, a single product). Thus, it's not applicable to records which assessment or evaluate more than one entities.

Sentence Level: This level task goes to the sentences and determines whether or not each and every sentence expressed a positive, negative, or neutral opinion. Neutral typically means no opinion. This sentence level of analysis is carefully related to subjectivity classification, which distinguishes sentences that express factual knowledge from sentences (referred to as subjective sentences) that specific subjective views and opinions.

Entity and Aspect level: Both the document level and the sentence level analyses do not discover what exactly people liked and did not like. Aspect level performs finer-grained analysis. Aspect level was earlier called feature level. Aspect level directly looks at the opinion itself. For example, although the sentence “although the service is not that great, I still love this restaurant” clearly has a positive tone, we cannot say that this sentence is entirely positive. In fact, the sentence is positive about the restaurant (emphasized), but negative about its service (not emphasized). In many applications, opinion targets are described by entities and/or their different aspects. Thus, the goal of this level of analysis is to discover sentiments on entities and/or their aspects.

Twitter: Twitter is a micro blogging website online that permit person to ship their tweet with the maximum characters used are one hundred forty characters. There are quite a few contents in Twitter corresponding to, Profile, Following, Follower, Mentions (@), Direct Message, Hash tags, Trending issues, and many others.

‘Tweets’ are seen by those who ‘follow’ the person who ‘tweeted’. Due to the growing popularity of the website,

Twitter can provide a rich bank of data in the form of harvested "tweets". Twitter by its very nature, allows people to convey their opinions and thoughts openly about whatever topic, discussion point or product that they are interested in sharing their opinions about. Therefore Twitter is a good medium to search for potentially interesting trends regarding prominent topics in the news or popular culture.

2. LITERATURE REVIEW

Analysis of Twitter information has been the focus of many contemporary researches within the domain of sentiment analysis. Many researchers are seeking to combine the sentiment evaluation and text mining as subsequent generation discipline. Level classification is most promising subject in sentiment evaluation document in Sentiment classification. Reference [11] showed that there is a correlation between sentiment measures computed utilizing phrase frequencies in tweets and both client self assurance polls and political polls. Accordingly, they illustrated that inclination of public towards special entities might be examined through analysis of tweets. Reference [12] measured presidential efficiency over a exact time interval by way of extracting general public sentiment from Twitter. For this motive they used the SentiStrength lexicon [13].As already acknowledged, [14] adopts a suite of sentiment features as well as some non-sentiment facets to procedure and analyze a manually annotated data set of tweets.

Many of the already existing PD systems nevertheless, participate in polarity detection without even defining a target phrase that their SA is directed at. In many-actual world issues, especially within the domain of purchaser products, assessment between a target item and its competitors will have to be handled. Hence, performing a target-oriented SA is important. Nevertheless, a unigram-situated model would become aware of the equal sentiment polarity for each goal of SA, as well as, its rivals that arise within the identical document. On this paper we reward a novel set of sentiment features to participate in goal-oriented PD on Twitter knowledge. We show that even by utilizing a very small set of elements, the unigram model (which is on the whole considered as a baseline for sentiment polarity analysis) is outperformed.

Many researchers have developed distinctive methods for sentimental analysis. The researcher Seyed-Ali Bahrainian et al. [9] presented a novel approach to SA of quick informal texts with a primary focus point on Twitter posts referred to as "tweets". He also compares state-of-the art SA procedure towards a novel hybrid process. The hybrid process utilizes a sentiment lexicon to generate a new set of features to instruct a linear support vector machine classifier. [15] awarded an adaptive sentiment analysis method called S-PLSA+, which no longer most effective can capture the hidden sentiment factors within the stories, but has the talents to be incrementally up-to-date as extra information grow to be on hand. And in addition exhibit how the proposed S-PLSA model can be utilized to sales efficiency prediction utilising the ARSA model.

Additionally the process proposed by means of Noriaki Kawamaet al. In [15] the place "the hierarchical technique to sentiment analysis, identifies each an item and its score by means of dividing topics, which is mainly handled as one entity. [16] developed novel sentiment ontology to conduct context-sensitive sentiment evaluation of on-line opinion posts in stock markets. ZHU Nanli et.Al. [5] introduced a survey on the cutting-edge progress in sentiment evaluation, and makes an in-depth introduction of its research and application in industry and Blogsphere. [18] adopts a suite of sentiment aspects as well as some non-sentiment aspects to procedure and analyze a manually annotated information set of tweets. [19] measured presidential performance over a special time period by using extracting general public sentiment from Twitter. For this purpose they used the SentiStrength lexicon [11]. Reference [2] confirmed that there's a correlation between sentiment measures computed utilising word frequencies in tweets and both patron self assurance polls and political polls. Thus, they illustrated that inclination of public in the direction of one-of-a-kind entities might be examined by analysis of tweets.

3. HYBRID POLARITY DETECTION

This section describes target-oriented hybrid sentiment analysis system. It consists of three major modules, a pre-processing module, a lexicon-based sentiment feature generator module and finally machine learning module. In the following subsections, each module is elaborated.

3.1 Preprocessing Module

This module performs a number of preprocessing steps as follows;
@username is replaced with "ATUSER".
URLs are removed.
"#word" is replaced with "word".
Slangs (abbreviations) are replaced with their actual phrase equivalences.
The target (of sentiment) word is replaced by "TARGET"

3.2 Sentiment Feature Generator Module

This module starts with replacing slangs with their equivalences using a slang dictionary. To build this slang dictionary, we use SentiStrength lexicon[13]. In the second step this module uses the SentiStrength lexicon [13] to tag all words present in dictionary for each document with their corresponding sentiment scores.

Likewise, according to a list of emoticons resent in dictionary, it tags happy emoticons with a sentiment score of "+1" and sad ones with a score of "-1". Also it further, tags all intensifiers (e.g. finally) and diminishers (e.g. may) with their corresponding scores. Also, it tags negation words with "NEGATE". Finally, if a word did not belong to any of the mentioned categories in the dictionary, it tagged that with the score "0". Having all words in a document tagged by their score now, we handle occurrence of intensifiers, diminishers, and negations. Firstly, we intensify the strength of a word that appears after an intensifier words, by the score of that intensifier word. Similarly, in the case of diminishers, we weaken the strength of a word that appears after a diminisher word by the strength of that diminisher. Finally, for negations, we flip the polarity of the score of a word that appears after a negation word. Then we weaken the flipped sentiment score by 1. That is, if the flipped score is positive, we subtract it by 1 and if it negative we sum it by 1.

Our primary goal in features extracting is to capture sequence of sentiment relevant words that show a document sentiment change. Additional, we define some features that present the neighborhood of the target of sentiment which we defined as iPhone. Table below shows this feature set. f1 is an overall sentiment score for an entire document and in order to compute this feature, we aggregate the words scores according to the tagging mentioned in the Sentiment feature generator module. We define the decision threshold zero for classifying words. If the score of a word is less than zero that word is tagged as negative, and otherwise if the score is greater than and equal to zero it is tagged as positive.

3.3 Machine Learning Classifier

F1	overall sentiment score
f2	count of positive words
f3	count of negative words
f4	count of negation words
f5	count of negation words followed by a positive word
f6	count of negation words followed by a negative word
f7	inverse sentiment
f8	count of positive words followed by target
f9	count of negative words followed by target
f10	count of negation words followed by target
f11	count of positive words followed by a negative word
f12	count of negative words followed by a positive word
f13	count of target words followed by a positive word
f14	count of target words followed by a negative word

The machine learning module is a linear support vector machine that takes as input the feature set described in the previous table and according to that classifies the tweets to separate classes.

The feature f7 named inverse sentiment, uses search for patterns that most of the words in a document have the opposite polarity of the actual overall document polarity. For example, the words "if", "after", "before" and "until" are among this category of words present in a document, the algorithm that f7 uses, further analyzes that document to discover if the sentiment words used in that document are meant to convey their opposite polarity. Now we explain the heuristic regarding "if statements" in detail. In the following example sentence is given

"If I don't get an iPhone for Diwali, I would be sad."

In the above statement, it shown that the actual polarity of the given sentence regarding target i.e. iPhone is positive, whereas, the sentence only contains words and patterns that usually occur in a negative context. "don't get an iPhone" is a negative sentiment regarding iPhone and it sets f10 to "1". Likewise, the word "sad" is a negative word and its occurrence sets f3 and f14 to "1". These features are often more likely to be present in negative sentences. However, because of the presenting patterns, the inverse sentiment feature (f7) is also set to "1". The heuristics search to find out if both the "if clause" and the "main clause" in the example sentence, include a negative sentiment pattern and if so it detects this pattern as an inverse sentiment. Therefore, our hybrid system can detect and correctly classify most such cases, whereas for many existing sentiment analysis systems handling such cases is a challenge.

4. EVALUATION

4.1 Dataset

Our dataset consists of 15,000 tweets. We gathered our dataset by consulting the Twitter API and making use of word spotting based on occurrence of the word "iPhone". We have no annotated data and hence we download tweets from twitter and classified with svm and MaxEntropy classifier. Tweets are classified as shown in following graph. The data classification with ManEntrpy gives better result as compared to svm.

4.2 Evaluation Metrics

We evaluate the methods presented in this paper using accuracy on overall accuracy as presented in the following:

Overall accuracy = (TP+TN)/ (TP+FP+TN+FN)

Where TP, FP, TN, and FN are the number of true positives, false positives, true negatives and false negatives. Furthermore, for testing any of the supervised classifiers as well as our hybrid method we use 10-fold cross validation.

4.3 Experimental Results

In this subsection we present the results of our experiments. Firstly, we test the data using two classifiers. If we see in below figure by classifying data using SVM and MaxEnt for 15000 tweets it having accuracy 88% and 87% with features respectively. Also for 5000 tweets it having accuracy 86% and 84% with features respectively.

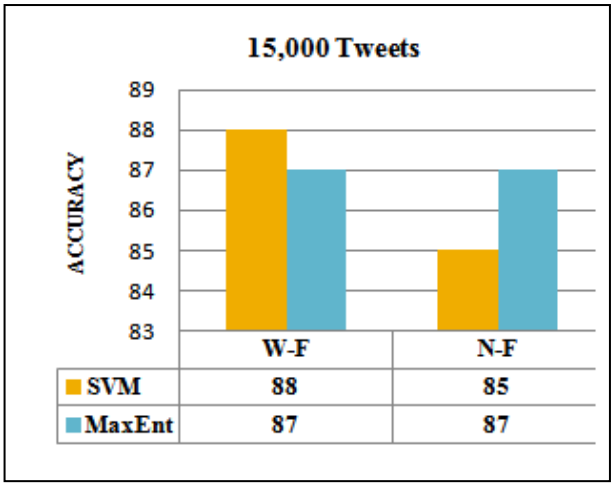

Chart -1: Classification of 15,000 Tweets with and without features

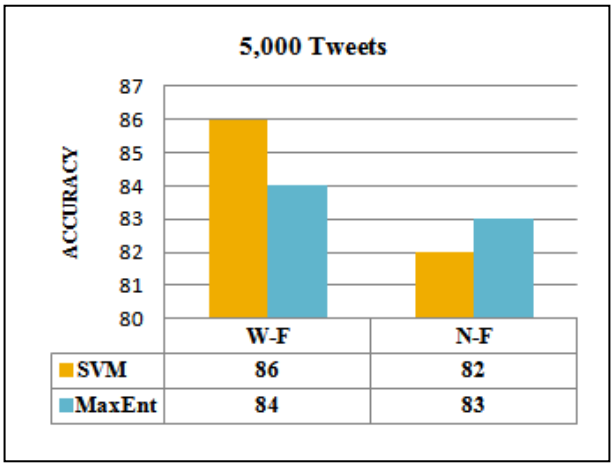

Chart -2: Classification of 5,000 Tweets with and without features

As shown in above figure, we conclude that using features for training classifier give maximum result as compared to without features. Also SVM give good result as compared to MaxEnt.

5. CONCLUSIONS

On this paper we introduces Hybrid method in which we combines Sentiment Lexicon with machine learning classifier for polarity detection of sentiment tweets in the area of social media. We conclude that according to our experiments, moving towards sentiment features rather than conventional text processing features would also be a promising solution to sentiment analysis. Discovering more points for sentiment evaluation which classify sentiments extra accurately is the future work of our study.

REFERENCES

[1] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai and A. Kappas, Sentiment strength detection in short informal text, Journal of the American Society for Information Science and Technology (2010), 2544–2558.

[2] Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.

[3] Tan,S., Wang, Y., & Cheng,X.(2008). "Combining learnbased and lexicon-based techniques for sentiment detection without using labeled examples", In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, July 20-24, 2008, Singapore, Singapore

[4] Pang, B., Lee, L., & Vithyanathan, S. (2002). "Thumbs Up? SentimentClassification Using Machine Learning Techniques." Proceedings of The ACL-02 conference on Empirical methods in natural language processing (pp. 79-86). Stroudsburg: Association for ComputationalLinguistic.

[5] Ley, Z., Riddhiman, G., Mohamed, D., Meichun, H., & Bing, L. (2011). "Combining lexicon-based and learningbased methods for twitter sentiment analysis". HP Laboratories, Technical Report HPL-2011, 89.

[6] B. O'Connor, R. Balasubramanyan, B. Routledge and N. Smith, From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series, in: International AAAI Conference on Weblogs and Social Media, North America, May 2010.

[7] M. Gamon, A. Aue, S. Corston-Oliver and E. Ringger, Pulse: mining customer opinions from free text, in: Proc. of the 6th International Conference on Advances in Intelligent Data Analysis, Madrid, Spain, September 8–10, 2005, pp. 121–132.

[8] H. Tang, S. Tan and X.A. Cheng, Survey on sentiment detection of reviews, Expert Systems with Applications: An International Journal 36(7) (2009), 10760–10773.

[9] S.-A. Bahrainian and A. Dengel, in: 2013 IEEE/WIC/ACM International Joint Conferences on Sentiment Analysis

Using Sentiment Features, Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Vol. 3, 17–20 Nov. 2013, pp. 26–29.

[10] S.A. Bahrainian, S.M. Bahrainian, M. Salarinasab and A. Dengel, Implementation of an Intelligent Product Recommender System in an e-Store, in: Proc. of the 6th International Conference on Active Media Technology (AMT'10), Toronto, Canada, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 174–182.

[11] O'CONNOR, B.; BALASUBRAMANYAN, R.; ROUTLEDGE, B.; SMITH, N.. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. International AAAI Conference on Weblogs and Social Media, North America, may. 2010.

[12] Lai, P., Extracting Strong Sentiment Trends from Twitter, 2011.

[13] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas. A., Sentiment strength detection in short informal text. Journal of the

[14] American Society for Information Science and Technology, pages 2544- 2558, 2010.

[15] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R., Sentiment analysis of Twitter data. In Proceedings of the Workshop on Languages in Social Media (LSM '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 30-38,2011.

[16] Xiaohui Yu,Yang Liu, Aijun An" An Adaptive Model for Probabilistic Sentiment Analysis", IEEE Computer Society ,Volume, Issue No. : 4191-4/10, pp-661-667, November 2010.