ISSN(Online): 2320 - 9801
ISSN (Print) : 2320 - 9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 6, June 2016

Sentiment Analysis of Tweets using Sentiment Features

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

M.Tech Student, Dept. of Computer Science and Information Technology, Dr B. A. M. University, Aurangabad, Maharashtra, India. [1]

Assistant Professor, Dept. of Computer Science and Information Technology, Dr B. A. M. University, Aurangabad, Maharashtra, India. [2]

ABSTRACT: A huge range of casual messages are announcing daily in informal community locales, net journals and exchange discussions. The blogosphere provides an expensive wellspring of knowledge regarding things, identities, innovations, and so on. Twitter could be a current continuous little scale blogging administration that allows its users to share short bits of knowledge referred to as "tweets". Users compose tweets to specific their feelings regarding completely different themes concerning their day by day lives. Actually, organizations collection such things have begun to survey these miniaturized scale websites to urge a sense of general opinion for his or her item. Usually these organizations study consumer responses and answer to shoppers on smaller scale sites. This paper introduces methodology utilizes a Sentiment Lexicon to return up with a group of choices to train a linear Support Vector Machine classifier.

KEYWORDS*:* Sentiment Analysis, Polarity Detection, Support Vector Machine, Opinion Mining, Machine Learning, Maximum Entropy, Naive Bayes, Sentiment Lexicon.

I. INTRODUCTION

Sentiment Analysis is the computational investigation of individual's conclusions, states of mind and feelings toward an entity. The entity can represent to people, occasions or themes. Opinion mining concentrates and investigates individual's opinion around an entity while Sentiment Analysis identifies the sentiment expressed in a text then analyses it. In this manner, the objective of sentiment analysis is to discover opinion, recognize the sentiment they express, and after that classify their polarity [1].

There are four levels of sentiment evaluation - Document level, Sentence level, Phrase level and Word level. In sentence, polarity categorization that aims to categories positive and negative sentiment for each and every tweets. Phrase level categorization may additionally be nested among sentence stage classification thus on capture multiple sentiments which can be reward among single sentence but the accuracy of predicting the sentiment is simply not acceptable. Consequently the latest method comes into image. a way supported on sentence as being subjective and so performs the sentiment classification on the subjective element of sentence. Most troublesome and powerful technique for sentiment classification that is predicated on document level that having 2 approaches: Term investigation and machine learning approach. Term counting approach entails deriving a sentiment measure by using calculating the positive and negative terms [20].

With the increasing quality of social networking, blogging and micro blog websites, everyday a big quantity of informal subjective text statements area unit created to be had on-line. The information captured from these texts, can be utilized for scientific surveys from a social or political point of view [2]. Businesses and product homeowners who intention to ameliorate their products/services may strongly benefit from the made suggestions [3], [4]. As an alternatively, customers may in addition taught about positivity or negativity of specific facets of products/offerings in line with users' opinions, to make an knowledgeable buy. In addition, purposes like rating on online movies based on online movie reviews [5] could not emerge without making use of these data.

Sentiment analysis is the computational system for extracting, classifying, figuring out and making a choice on the opinions expressed in various contents. Researchers have additionally recognized the have an impact on of online experiences; and have produced some essential outcome on this field. A couple of internet sites furnish facilities for the users to publish their opinions in their websites equivalent to Amazon. Reviews are viewed to be centered in blog posts, social networking websites as well as dedicated evaluation websites similar to opinions. This online evaluate is considered to be a priceless instrument for firms. Hence, on-line reviews can also be very valuable, as at the same time such stories replicate the "knowledge of crowds" and is usually a excellent indicator of the product's future income efficiency [6] ,[7].

Twitter messages as several others announce on the blogosphere having lot of informal texts. Since of the anomalistic nature of casual text, evaluation or processing of this variety of textual content is probably tougher if when put next with formal texts. The most important difference between processing formal textual and informal textual content is in information pre-processing. Formal textual content commonly desires less pre-processing. Informal textual content nevertheless, generally includes emoticons, use of slangs, bad grammar, and sarcasm or non dictionary commonplace phrases. As a result, analysis of this category of text is regularly more elaborate. The foremost purpose of this work is processing informal text statements.

We first test the consequences of presence or absence of emoticons and discontinue words in a unigram characteristic vector to define a baseline. We train the two classifiers, support vector machine and naive bayes centered on the unigram function set [5] and compare them against our hybrid process.

Polarity Detection : Polarity Detection is most likely referred to as classifying subjective opinions into Positive and Negative. Present polarity detection ways classified into three predominant classes: supervised, unsupervised, and hybrid.

Supervised Methods: The supervised finding out approaches rely upon the existence of labelled training data. Supervised approaches are Machine Learning (ML) methods in which a classifier is trained based on a feature set, utilizing labeled training data. Pang et.Al. [6] have been one of the very first to perform SA on online film reports. They established ML classifier, specifically, SVM, MaxEnt and NB classifiers, and trained them on unigram of bag of word. Their findings confirmed that an SVM trained on a unigram bag-of-words characteristic set, outperforms all different techniques presented of their work. Drawbacks of classifier are that for polarity detection they are both domain (dataset subject) and temporally (datasets collected in distinctive time durations on an annual groundwork) dependent [8]. Additionally, for classifiers to be informed on significant function sets (e.g. Unigrams) long training sessions are required and as a result of the tremendous quantity of features they have a giant time and reminiscence complexity.

2. Unsupervised Methods: The unsupervised method are used when it is difficult to find these labeled training data. Unsupervised methods are more often cantered on a Sentiment Lexicon, where each and every sentiment-bearing word is associated with both a sentiment rating or a collection of sentiment bearing seed phrases. These methods use different algorithms to compute a sentiment for a given document. A SL might be mechanically generated utilizing a suite of positive and negative seed-words [9], [10] or it could be manually built. There exists a quantity of such manually developed senti strength lexicon. The Senti strength lexicon [11] , which is specifically designed for casual texts. Drawbacks of lexicon-based ways are domain dependency, effort for obtaining a lexicon, as well as retaining them due to language change over time.

3. Hybrid Methods: Hybrid method undertakes a combination of each of the above mentioned classes to perform opinion mining [12]. A sentiment lexicon is used to generate points for training an machine learning classifier. Hybrid method would resolve many of the major issues of supervised method, specifically, the difficulty of huge characteristic units and time and reminiscence complexity. On the opposite, a necessity for area stylish lexicons nonetheless exists. Hybrid process combines utilization of a sentiment lexicon along with a machine learning classifier for polarity detection of opinionated texts within the domain of client-products. A goal–oriented hybrid procedure outperforms the unigram baseline. All the features proposed within the hybrid system, require a very quick time to be computed.

II. RELATED WORK

There are a couple of ways proposed for sentiment analysis for Twitter data. Evaluation of Twitter knowledge focuses of many recent researches within the domain of sentiment analysis. Many researchers have developed distinctive methods for sentimental analysis. The researcher Seyed-Ali Bahrainian et al. [13] 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. [14] 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. This technique integrates preferred sentiment analysis into computing device finding out strategies situated on aid vector laptop. 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 Blogosphere. [18] adopts a suite of sentiment aspects as well as some non-sentiment facets to procedure and analyse 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.

Hybrid Polarity Detection System: Hybrid polarity detection system consists of three module: a pre-processing module, a lexicon-based sentiment function generator module and machine learning module.

A. The Pre-processing Module: This module performs a number of pre-processing steps as following;
1.@username is replaced with “ATUSER”.
2. URLs are eliminated.
3. “#word” is replaced with “word”.
4. Slangs (abbreviations) are changed with their specific phrase equivalences. A manually constructed slang dictionary is used for this reason [11].
5. The target word is replaced by “TARGET”.

B. Sentiment Feature Generator Module: This module begins with changing slangs with their equivalences utilizing a slang dictionary. To construct this slang dictionary, we manually accumulated an awfully complete slang dictionary by means of using as many online resources that we might find. Then, within the 2nd step this module makes use of the Senti strength lexicon [11] to tag all sentiment-bearing phrases in every document with their corresponding sentiment ratings. Likewise, consistent with a record of emoticons, it tags joyful emoticons with a sentiment score of “+1” and unhappy ones with a score of “-1”. It additional, tags all intensifiers (e.g. finally) and diminishers (e.g. may) with their corresponding rankings.

Additionally, it tags negation phrases with “NEGATE”. If a word did not belong to any of the mentioned classes, it will be tagged with the ranking “zero”. After having all phrases in a file tagged both by means of their score or type, now we will have to handle incidence of intensifiers, diminishers, and negations. First, we intensify the strength of a sentiment-bearing phrase that appears after an intensifier, by way of the ranking of that intensifier phrase. Analogously, in the case of diminishers, we weaken the strength of a sentiment-bearing word that appears after a diminisher phrase with the aid of the strength of that diminisher. In the end, for handling negations, we flip the polarity of the score of a sentiment-bearing phrase that seems after a negation. Then we weaken the flipped sentiment rating by way of 1. That's, if the flipped ranking is positive, we subtract it by means of 1and if it’s negative we sum it by means of 1. Note that in all instances we ignore the “zero” tags that show up in between probably the most above mentioned valence shifters and a sentiment bearing word in a single text snippet, at the same time performing the above stated computations.

Feature Selection: Our primary goal in extracting features is to capture sequence of sentiment words that shows sentiment change. Following table offers this selection set. Characteristic f1 is an total sentiment rating for an entire tweet. With the intention to compute this feature, we mixture the words scores consistent with the tagging system. We outline the decision threshold of ‘0’ for classifying words. That's, if the sentiment score of a word is lower than 0 that word is tagged as negative, and in any other case if the rating is greater than 0 it's tagged as positive.

Table 1: Features used for sentiment Analysis

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	count 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

C. Machine Learning Classifier

The machine learning module is a linear support vector machine that takes input as characteristic set described in the earlier subsection and thus classifies the tweets to separate courses.

III. DATASET

Dataset consist of 940 positive tweets and 940 negative tweets. We collect our tweets by consulting the Twitter API and making use of word spotting based on occurrence of the word “iPhone”.

IV. EXPERIMENTAL RESULT

We compare SVM as baseline and our Hybrid method we get 82.32% and 85.03% accuracy respectively. In Hybrid method we have 14 features and linear SVM takes as input these features and accordingly classify tweets. All these features in hybrid system require very short time to be computed.

Table 2: Comparison between Hybrid approach and SVM

	Overall Accuracy(%)
Hybrid Approach	85.03
SVM as baseline	82.32

DOI: 10.15680/IJIRCCE.2016. 0406078

V. CONCLUSION AND FUTURE WORK

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 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. Tsytsarau Mikalai, Palpanas Themis. Survey on mining subjective data on the web. Data Min Knowl Discov 2012;24:478–514.
2. 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.
3. Gamon, M., Aue A., Corston-Oliver, S., Ringger, E., Pulse: mining customer opinions from free text, Proceedings of the 6th international conference on Advances in Intelligent Data Analysis, p.121-132, September 08-10, 2005, Madrid, Spain.
4. Tang, H., Tan, S., Cheng, X., A survey on sentiment detection of reviews, Expert Systems with Applications: An International Journal,v.36 n.7, p.10760-10773, September, 2009.
5. Pang, B., Lee, L., and Vaithyanathan, S., Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10 (EMNLP '02), Vol. 10. Association for Computational Linguistics, Stroudsburg, PA, USA, 79-86, 2002.
6. 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.
7. Pimenta, F, Obradovic, D., Schirru, R., Baumann, S., Dengel, A., "Automatic Sentiment Monitoring of Specific Topics in the Blogosphere", Workshop on Dynamic Networks and Knowledge Discovery (DyNaK 2010), Barcelona, Spain, September 2010.
8. Read, J., Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In Proceedings of the ACL Student Research Workshop, 2005.
9. 9 Pimenta, F, Obradovic, D., Schirru, R., Baumann, S., Dengel, A., "Automatic Sentiment Monitoring of Specific Topics in the Blogosphere", Workshop on Dynamic Networks and Knowledge Discovery (DyNaK 2010), Barcelona, Spain, September 2010.
10. Turney, P. D. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 417–42,2002.
11. Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas. A., Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, pages 2544-2558, 2010.
12. Diana Maynard, Adam Funk. Automatic detection of political opinions in tweets. In: Proceedings of the 8th international conference on the semantic web, ESWC'11; 2011. p. 88–99.
13. Seyed-Ali Bahrainian ,Andreas Dengel, "Sentiment Analysis using Sentiment Features" ,IEEE Computer Society ,Volume, Issue No. : 2902-3/13, pp-26-29, 2013.
14. 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.
15. Noriaki Kawamae"Hierarchical Approach to Sentiment Analysis", IEEE Computer Society, Volume, Issue No. : 4859-3/12,pp- 138-145, 2012.
16. Desheng Das Wu, Lijuan Zheng ,and David L. Olson ,"A Decision Support Approach for Online Stock Forum Sentiment Analysis", IEEE TRANSACTIOND ON SYSTEMS, MAN ,AND CYBERNETICS SYSTEMS, vol-44,issue no-8, pp:1077-1087, August 2014.
17. ZHU Nanli,, ZOU Ping, LI Weiguo, CHENG Meng "Sentiment Analysis: A Literature Review", proceedings of IEEE, 2012.
18. 18. 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.
19. Lai, P., Extracting Strong Sentiment Trends from Twitter, 2011.
20. Aamera Z. H. Khan, 2Dr. Mohamma Atique, Dr. V. M. Thakare, "Sentiment Analysis Using Support Vector Machine". International Journal of Advanced Research in Computer Science and Software Engineering,Volume 5, Issue 4, April 2015.