Volume 5, Issue 7, July 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Present New Sentiment Analysis Framework in Twitter

[1]Fatemeh Forouzesh*, [2]Ahmad Farahhi

[1]M.Sc Student of Computer Engineering (Software), Payame Noor University, PO BOX 19395-3697, Tehran, Iran

[2]Assistant Professor of Department of Computer Engineering and Information Technology,
Payame Noor University, Tehran, Iran

Abstract— The emergence of Web 2.0 has drastically altered the way users perceive the Internet, by improving information sharing, collaboration and interoperability. Nowadays, Sentiment Analysis (SA) is receiving huge attention because of the wide range of its direct applications like analyses of products, customer profiles, and political trends. Twitter is one of the most important micro blogging that will help all individuals, businesses, organizations in sharing their ideas and opinions about special brand or issues. This site has nearly 600 million users and over 250 million messages per day as one of the most popular social networks are most important. Twitter sentiment analysis on data explores people feelings about the brand satisfaction degrees and offer business managers in order to present best services to customer in a quick and efficient way. In this paper introduces the Sentiment analysis methods and describe some problems and the strengths of the previous procedures to challenge and design new framework to enhance accuracy. Also we show the impact of pre-processing as first step in sentiment analysis.

Keywords— Sentiment analysis, Classification, social media mining, Opinion Mining

I. INTRODUCTION

Online discussions play a significant role mainly because of their content. Since people use this kind of discussions in order to express their opinions and exchange ideas .Social media platforms offer individuals the opportunity to articulate opinions on various topics ranging from consumer products and services to social political issues. These opinions are quite useful in various areas such as marketing for managers or policy making for government agencies. There are several names for sentiment analysis like Opinion Mining, Opining Extraction, Sentiment Mining, and Subject Analysis. In sentiment analysis, the classes to which a piece of text is assigned are usually negative or positive. There are some Twitter-specific sentiment analysis studies. Twitter sentiment analysis is a bit different from the general sentiment analysis studies because Twitter posts are short. The maximum number of characters that are allowed In Twitter is 140. Moreover Twitter messages are full of slang and misspellings [6]. Because of this mere fact, some classification approaches and feature selection used in general sentiment analysis may not be important to twitter sentiment analysis. In this paper we combined two techniques in sentiment analysis to evaluate our model. And show that the pre-processing step is really important to improve accuracy. The following are some challenges faced in Sentiment analysis pre-processing of Twitter feeds.

- Remove waste and extra characters such statements? @ ^ #.
- Considering the negation instruments to determine the exact characteristics polarity.
- Check the correct pronunciation of words and correct them automatically.
- Use common terms dictionary to replace the word Twitter stands.
- Use emotional dictionary to determine the polarity of each Tweet.
- Combine lexicon-base technique wit machine learning.
- Compare our framework with other supervised technique.

II. LITERATURE REVIEW

A lot of work has been done in the field of opinion mining or sentiment analysis for well over a decade now. Different techniques are used to classify the text according to polarity. Most of these techniques can be classified under two categories: Machine learning and Lexicon Based.

[2, 3, 4, 5, 14, 16, 17, 18] use Dictionary-based method for the classification of documents to determine the degree of polarization of each document. For compute the polarity of each document they use specific function.

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity. Starting from being a document level classification task [15, 16, and 22], it has been handled at the sentence level [1, 13] and more recently at the phrase level [13] the proposed research is in the area of sentiment analysis. To determine whether a document or a sentence expresses a positive or negative sentiment, two main approaches are commonly used: the lexicon-based approach and the machine learning-based approach. Popularity of lexicon-based approaches is rapidly increasing since they require no training data, and hence are more suited to a wider range of domains than supervised approaches. Nevertheless, because of the lexicon-based approaches have the number of words in the finite lexicons has a major limitation. [1, 13] show that the lexicon-based approach determines the sentiment or polarity of opinion via some function of opinion words in the document or the sentence. For entity-level sentiment analysis method has low recall.

The machine learning-based approach typically trains sentiment classifiers using features such as unigrams or bigrams [23]. Most techniques using a supervised learning using various learning techniques such as Naive-Bayes, maximum entropy and support vector machines. For each application domain these methods needs manually labelled.
There are also some approaches that utilizes both the opinion words/lexicon and the learning approach. For example, [20] used a subjectivity lexicon to identify training data for supervised learning for subjectivity classification. Our work does not do subjectivity classification. A similar idea was also applied to sentiment classification of reviews in [9], which classifies reviews into two classes, positive and negative, but no neutral class, which makes the problem much easier. While most sentiment analysis methods were proposed for large opinionated documents (e.g. reviews, blogs), some recent work has addressed micro blogs [19] proposed a two- step classification method. There are also some approaches that utilizes both the opinion words/lexicon and the learning approach. For example [20] used a subjectivity lexicon to identify training data for supervised learning for subjectivity classification.
In this paper we propose the new hybrid framework that uses two dictionaries to calculate polarity for each tweet .one dictionary has positive and negative word polarity and the other one has character polarity. Our method is similar to [21]

But in two aspects has difference. Firstly He just used word polarity dictionary to determine the polarity of each sentences but in our work we have two dictionary one for word and the other for emotion character. We believe that by using this dictionary we can improve accuracy of determine polarity. Secondly in our method we get the polarity score by propose algorithm and then we use this labelled sentences as training input for learning machine. Which means we don't use labelled manually data.

III. SENTIMENT ANALYSIS TECHNIQUES

Generally speaking‹sentiment analysis aims to determine the attitude of a speaker or writer with respect to some topic or the overall contextual polarity of a document. Micro blogs such as Twitter have become a rich and valuable source of information. This is due to their nature that includes posts of users in real time about various topics of common issues; complaints and expression of public emotion in daily life. The manufacturers of a product explore these comments to elicit public emotion on a product or a brand.

A. Unsupervised Learning Techniques

Unsupervised learning is base on Lexicon Techniques that work on an assumption that the Collective polarity of a document or sentence is the sum of polarities of the individual words or phrases. [25]In this technique, classification is performed based on a function that compares characteristics of the text to the words that their polarity is already marked. For example, at first, it takes the positive and negative words out from the text. Then, it counts number of positive and negative words. If the number of positive words was more than negative words the polarity of sentence in considered as positive. Otherwise, the polarity is negative. A similar achievement was proposed by [22]. They used an unsupervised technique for criticisms classification as suggestions. He used this technique to determine the polarity of the words used. He calculated in each criticism the average polarity of words. Then, according to the mean, he divided criticisms into two categories: positive and negative.

B. Supervised Learning Techniques

Supervised learning based on Machine learning strategies that work by training dataset. That is, a machine learning algorithm needs to be trained first for both supervised and unsupervised learning tasks. [12] The main task of supervised techniques is building a classifier. Classifier needs the examples for training which can be labelled manually or by the manufacturer form online resources. Supervised algorithms that are mainly used are: Support Vector Machine (SVM), Naïve-Bayes and Multinomial Naïve-Bayes. [15] Showed the supervised techniques versus unsupervised technique are more efficient. Supervised techniques can be obtained from one or a combination of algorithms. For supervised techniques, the analysed text should be given as a vector of features. Used features in the vector are one or a combination of the features.

IV. PROPOSED APPROACH

A background study reveals that the process of opinion target extraction involves various natural language processing tasks and pre-processing techniques such as: tokenization, part-of-speech tagging, noise removal, feature selection and classification.

The proposed approach is combination of learning machine technique and lexicon based technique i.e. a dictionary of sentiment bearing words along with their polarities was used to classify the text into positive, negative or neutral opinion .our propose framework has three phases. The first phase is pre-processing step. This phase involves removing characters or sequences of characters that cannot assist during the subsequent sentiment analysis phase, in order to reduce the noise in the data set. The second phase is determining the polarity of each tweet by two dictionary lexicon .and the third phase is applying Labelled data as trained dataset for learning machine. Finally we tokenized the dataset and constructed n-grams. Then we experimented with several classifiers including SVM, Naive-Bayes and K-NN to reach best result .These phases are showed in "Figure 1. Propose framework",

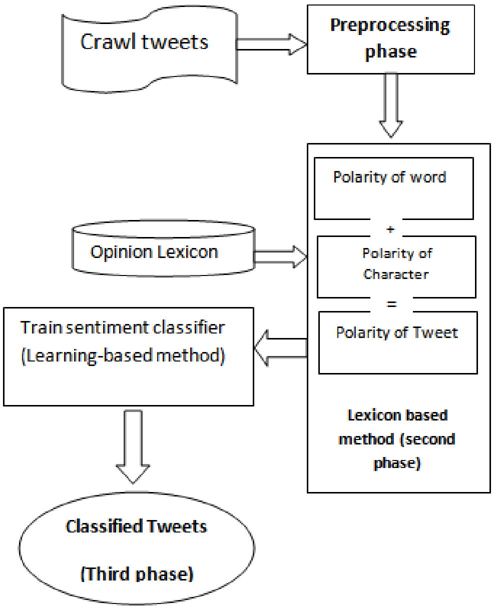

Fig. 1 three phases in our framework for sentiment analysis

V. DATASET

One way to collect data set is using third party site. For the work and experiments described in this paper; we used example data set that collected from Zapier[1] site. This third party site allow users to create a new zap application to collect data from twitter in excel sheet. In order to search from twitter we need twitter account for registration. This dataset consists of 2000 tweets opinion randomly selected about iphone mobile. 700 tweets in this dataset contain positive emoticons, such as :), :-), :), :D, and =), the 700 contains negative emoticons such as :(, :-(, or: (and 600 tweets such as :O for neutral.

VI. SENTIMENT CLASSIFICATION

As all people from every country can share their opinion in twitter, our dataset has many tweets from other languages like Arabic, France, Germany. Detect Language was the first method on data from the Twitter campaign. In this paper we just use English tweets and regard other language tweets.

In pre-processing phase we cleaned the data by removing URL links, user names (those that are marked by @), RT (for re-tweet), the emoticons, and stop words.

In our framework we used two new resources for training twitter data: 1) an emoticon dictionary 2) an acronym dictionary[2] which contains 5298 words. We prepare the emoticon dictionary by labelling 170 emoticons listed on Wikipedia[3] with their emotional state. For example, “:=)” is labelled as positive whereas “:=(” is labelled as negative and “:O” is labelled as neutral . We assign each emoticon a label from the following set of scores from -5 to +5.note that Zero is assign for neutral tweets. We change acronym dictionary word to correct word. Some acronyms are showing in “Table 1. Acronym Example”. Then remove special character like @, #,*, ^, $ from tweets. For this we use regular expression In order to eliminate redundant words. And also we remove stop English word from each tweet.

Table I. Example of acronym and expressions are equivalent in dictionary.

Acronym	English Expansion
TYVW	Thank you very much
YT	You Tube
JSYK	Just so you know
ICU	I see you

Negation word like don’t,didn’t,not,… play a key role in determining the polarity of tweets. Negation reverses the meaning of phrase. [2] Performed sentiment analysis while handling negation and intensifying words unfortunately that is regardless in the previous article. Therefore we check if the negation word appears before adjective, our application reverse the score of adjective by multiply at -1. Although the word ‘good’ depicts a positive sentiment the negation – ‘not’ reverses its polarity. In the proposed approach whenever a negation word is encountered in a tweet, its polarity is reversed.

[1] www.zapier.com
[2] http://www.noslang.com/
[3] www.allacronyms.com/twitter/topic

After this step, we check the spell of word and automatically correct misspelling phrases with spell checker application. For example the word “gooood” should replace with “good”. For Sentiment analysis not only phrases and words are important but also sometimes emotional character of tweets can expressed the feelings of people. Some emotional characters are show in “Table 2. Emotion faces”

Table III. An example of positive and negative emotional faces

Happy emotional characters	**Sad emotional characters**
:p :}	:D }=
:-) :--}	:(=={
:<) :-3	:-{ :,(
:-] ^_^	:(:-{{

The algorithm that we use for calculate polarity is as describe follow. Note that in this code we merge to polarity dictionary and Emotional dictionary with together. This algorithm is similar to[23] which used blind negation to determine polarity.

Input: Tweets, SentiWord Dictionary, Negation List, Emotion Dictionary, Output: Sentiment (positive, negative or neutral)

```
BEGIN
For each tweet Ti
{
Score = 0;
For each word Wj in Ti that exists in Sentiword Dictionary && Emotion Dictionary
{
If polarity [Wj] = positive|| negative && Wj-1!=NegationList
{
Score = Score + polarityDIct[Wj];
}
Else If
{Score = Score + polarityDIct[Wj]*-1;
}
If score of Ti >0
{
Sentiment = positive
}
Else If Score of TI<0
{
Sentiment = negative
}
Else
{
Sentiment = neutral
}
Return Sentiment
}
END
```

which means that if the sum of positive polarity Tweet greater than zero the tweet is consider as positive, if less than zero its negative or if zero the tweet is neutral.[22] used similar procedure to classify your document.

The main difference the work done by their work with our framework presented in this paper is preprocessing step that we use to improve accuracy. He's also done his work on the document level but we do it at sentence level.

In order to score the tweet we search each word and symbol in two dictionaries. For lustrate our algorithm we describe an example.

“ My iphone voice call is very good :)But I don’t like its color because it’s ugly”.

After applying the steps outlined above, the rate is as follows. The polarity of each word and character is showing in “Table 3. Polarity of example sentence”

Table IIIIVI. Polarity of example sentence

good	+3
(:	+3
like	+2
don’t	-2

like	
ugly	-3

Sum of polarity in this example is: +3+3-2-3=+1
As 1>0 so Polarity=positive
According to below function:

$$\text{Polarity (d)} = \begin{cases} \text{Positive} & \sum_{t \in T(d)} o(t) > 0 \\ \text{Negative} & \sum_{t \in T(d)} o(t) < 0 \\ \text{Neutral} & \sum_{t \in T(d)} o(t) = 0 \end{cases}$$

Since the rating of the tweets was positive as a result of the following formula we consider it as a positive tweet.

The bag-of-words is used in several approaches for document classification and named entity extraction. This feature incorporates the use and frequency of individual words or phrases, and it disregards contextual and syntactic relations of words in sentences or documents. Term frequency inverse document frequency (TF-IDF) models have exploited the bag-of-words representation for document classification. The bag-of-words feature has also been effectively employed for opinion target and sentiment extraction [24, 26] in third phase we use TF-IDF that Generates word vectors from string attributes for classification in learning machines.

Here is how the keyword-based sentiment classification algorithm works: for each tweet, the number of positive and negative keywords and the emotional faces found in it is counted; this is done by cross referencing the words in the tweet with the respective keyword lexicon and emotional faces lexicon. Every word and emotional faces has the score .we sum the score of each tweet .If a tweet has positive score then it is a positive tweet; if it has negative score, then it is a negative tweet; if it has zero score contains, which means the number of score in positive and negative is equal it is neutral . This classifier was run on the test data. The test data contains 2000 tweets of which 700 are negative, 700 are positive and 600 are neutral. This data is the standard test data. The results are presented in “Table 4. Keyword-based classification on test data” in the form of a confusion matrix. In the confusion matrix, the sum of the cells in a row makes the total of the actual class. For example, total number of positives is 115+525+60. The number at the intersection of positive column and positive row is the number that is correctly classified as positive. For positive, 525 out of 700 are classified correctly. The rest are classified incorrectly as 60 negative115 neutral. Thus, the per class accuracies are:

accuracy_{neg} = 450/700 = 64.2%

accuracy_{pos} = 525/700 = 75%

accuracy_{neu} = 415/600= 69.1%

What the confusion matrix in and the accuracies show is that the classifier Performs low in recognizing negative tweets. However, it does very well in recognizing neutral tweets. The result on positive tweets is not that bad too, 55.8%. The overall accuracy is the percentage of tweets that are classified correctly. The sum of the three cells on the diagonal from top left to the bottom right is what is correctly classified.

Accuracy $_{total}$= 1390/2000 = 69.5%

Table IV. keyword-based classification on test data

negative	positive	neutral	
450	120	130	negative
60	525	115	positive
120	65	415	neutral

Table V. Result of sentiment analysis in different algorithms

Machine learning algorithm	accuracy
Naïve bayes	62.02%
SVM	64.53%
K-NN	52.71%
Propose Framework	**69.5%**

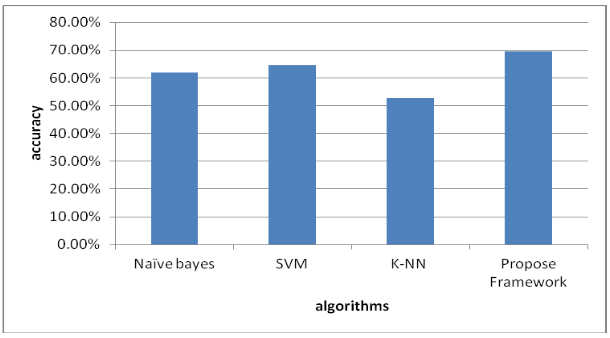

Fig 2. Comparison of the proposed approach with other methodologies

VII. RESULTS AND DISCUSSION

As we known review of people opinion can be very useful for business marketing. We get reaction of people for particular products, events, issues very fast on web. Sentiment analysis is used in different domains. For our propose framework we chose opinion about iphone mobile as dataset. And compare our framework with other existing techniques such as machine learning. Our frame work is a combination of lexicon based approached a supervised learning. The result shows that, in compare with other techniques our propose framework has better accuracy.

In future we will apply this approach for the other close domain. We would like to extend the results presented here by evaluating our methods in real time and using other languages tweets like Farsi. And also we plan to apply our proposed method to the personalized tweet recommender system based on the tweet history and social relations of users.

ACKNOWLEDGMENT

The authors would like to thank fellows of the international Technology conferences for their reviews on this paper. I’m grateful to my guide Prof & Head Farahhi for his valuable suggestions and encouragement. Special thanks to Mr. Dadfar for giving me constant guidance and suggestions.

REFERENCES

[1] S.Kim,E. Hovy, “Determining the Sentiment of Opinions”. COLING'04 Academy of Management proceedings ’96, 1, 137–141, 2004.

[2] M.Taboada, J.Brooke,M. Tofiloski, K.Voll, and M.Stede,” Lexicon-based Methods for Sentiment Analysis”. Journal of Computational Linguists, 2010.

[3] X.Ding, B.Liu, and P.Yu, “A Holistic Lexicon-based Approach to Opinion Mining”. WSDM 2008.

[4] N.Kaji, M.Kitsuregawa, “Building lexicon for sentiment analysis from massive collection of HTML” documents. In Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning, (EMNLP-CoNLL) (pp. 1075–1083), 2007.

[5] A.Moreo, M.Romero, J.Castro,J. Zurita, “Lexicon-based comments oriented news sentiment analyzer system”. Expert Systems with Applications 39(10), 9166–9180, 2012.

[6] S.Weiss,N.Indurkhya,T. Zhang, and F.Damerau, ,” Text mining Predictive methods for analyzing unstructured information”. Disciplinary fairness by unionized employees and disciplinary subject matter experts. Journal of applied psychology, 82, 699–705, 2005.

[7] S.Masterson, M.S .Taylor, “The broadening of procedural justice: should interactional and procedural components be separate theories?” Paper presented at annual meeting of the Academy of Management, Cincinnati, OH, 1996.

[8] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li.” Sentiment classification of Internet restaurant reviews written in Cantonese”, Expert Systems with Applications Songbo Tan, Jin Zhang, “An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications 34 (2008) 2622–2629, 2011.

[9] R.Prabowo, M.Thelwall.. “Sentiment analysis: A combined approach. Journal of Informetrics “3. 143–157,2009

[10] X.Kaiquan , S.Stephen,L. Jiexun, S.Yuxia,”.Mining comparative opinions from customer reviews for Competitive Intelligence”. Decision Support Systems 50 743–754, 1996.

[11] Ku.Lun-Wei, Lee, Chia-Ying, Chen, Hsin-Hsi, “Identification 1655 of opinion holders. Compute”. Linguist. Chin. Lang. Process. 14 (4), 1656 383–402. 2009.

[12] A.Aue, M.Gamon, M, “Customizing Sentiment Classifiers to New Domains: a Case Study”. RANLP, 2005.

[13] M.Hu, B.Liu, “Mining and summarizing customer reviews”. KDD'04, 2004.

[14] D.De Kok,H. Brouwer,” Natural language processing for the working programmer”. E-book available under the creative commons attribution 3.0 License (CC-BY), 2012.

[15] Pang, B., and Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

[16] A.Aue, and M.Gamon, “Customizing sentiment classifiers to new domains: a case study”. In Proceedings of the international conference on recent advances in natural language processing (RANLP-05) (pp. 207–218), 2005.

[17] S.Baccianella, A.Esuli, F.Sebastiani,” Sentiwordnet 3.0: An enhanced lexical 563 resource for sentiment analysis and opinion mining”, in: Proceedings of the 7th 564 International Conference on Language Resources and Evaluation, 2010, pp. 5652200–2204.

[18] E.Cambria,R.Speer,C.Havasi, A.Hussai” semantic resource for opinion mining. Senticnet”: A publicly available 579, in: AAAI Fall Symposium: 580, 2010.

[19] L.BARBOSA ANDJ. FENG, J,” Robust sentiment detection on twitter from biased and noisy data”. In Proceedings of COLING. pp. 36–44, 2010.

[20] J.Wiebe, and E. Rilofi ,”Creating Subjective and Objective Sentence Classifiers from Unannotated Texts”. CICLing, 2005.

[21] C.Kaushik , A.Mishra,” A SCALABLE, LEXICON BASED TECHNIQUE FOR ANALYSIS” Computer Engg. Department, YMCA University of Science & Technology, SENTIMENT Faridabad International Journal in Foundations of Computer Science & Technology (IJFCST), Vol.4, No.5, September 2014.

[22] Turney. “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews”. ACL, 2002.

[23] Pang and L. Lee., 2004. A sentimental education: Sentiment analysis using subjectivity analysis using subjectivity summarization based on minimum cuts. ACL.

[24] K.Nigam, M.Hurst, "Towards a Robust Metric Of1737 Opinion". Paper Presented At the AAAI Spring Symposium On1738 Exploring Attitude and Affect In Text, 2004.

[25] P.MELVILLE, W.GRYC, R.LAWRENCE,"SENTIMENT ANALYSIS OF BLOGS BY COMBINING LEXICAL KNOWLEDGE WITH TEXT CLASSIFICATION". KDD, 2009.

[26] L. and Feng, J." Robust Sentiment Detection on Twitter from Biased and Noisy Data",2010.