Pak. J. Biotechnol. Vol. 15 (Special Issue 1) Pp. 22-26 (2018) Sultana A.R. et al.,
www.pjbt.org, PISSN: 1812-1837, EISSN: 2312-7791

TEXT BASED SENTIMENT ANALYSIS OF PRODUCT REVIEWS

A Razia Sulthana, A Arokiaraj Jovith and L Sairamesh

Department of Information Technology, SRM University, Kattankulathur, India
Department of Information Science and Technology, Anna University Chennai, India
raziasulthana.a@ktr.srmuniv.ac.in, aarokiarajjovith.a@ktr.srmuniv.ac.in, sairamesh.ist@gmail.com

ABSTRACT

With advent of web 2.0 people started sharing their opinions in social network. The social media helps in communicating with public and provides a clear platform to share the views about the product. This has led to different ways of analyzing the user reviews. Sentiment analysis is one of the wide spread area which helps in identifying the same. We proposed a sentiment classifier which recognizes the opinion word based on linguistic analysis. This analysis is done before the preprocessing stage so as to filter outliers and extract only the necessary words. In against the existing approach, the time consumed in this method is considerably reduced as the reviews are analyzed in the initial stage. This method analyzes the sentiment of the reviews posted by customers in online portals by taking the bigrams into account. The relationship between bigrams is identified to know the wavelength of the user's intention. The prioritized bigrams are chosen for every review such that it qualifies the root word and the root word itself. To inculcate we have implemented a different theoretical model. The data set we have taken for our experiment is a collection of 25,000 reviews from Cornell. The model was experimented with different training sets where the accuracy and precision measures shows a marginal increase. The results of our approach can be used for predicting the results in future as per the market specifications and future models. The accuracy, precision and recall were the metrics that were used to identify the quality of our methodology. Thus the study on bigrams in reviews yields an added value in sentiment analysis.

Keywords: *Classification, K-Means Clustering, Ontology, Preprocessing, Review, Recommendations.*

1. INTRODUCTION

Svetlana et al., (2014) framed system to detect the sentiment of short tweets, short messages (message level task) and identifies the sentiment of a term within a message (term level task). This approach classifies the text in a supervised manner based on the semantic and sentiment features. The features are derived using high coverage tweet specific sentiment lexicons. It captures the negated words from the tweets using a distinct sentiment lexicon. This framework is implemented in SemEval-2013 SMS data set and a corpus of movie review excerpts and the performance measures were calculated. It is also experimental proven that the performance measure show a significant increase when the lexicons were automatically generated

Nathan et al., have described the sentiment of tweets classified using an algorithm based on the high- level features extracted from the tweets. The characteristics of the tweet are also identified to be either subjective or objective with minimal error rate. This method can be applied in environments where batch learning process is trivial. It is implemented in Sanders Corpus Data set which encapsulates 5513 hand classified tweets. The ensemble method used here classifies using multiple gram size which helps in approximate prediction of tweets. Many classification algorithms simultaneously predict the sentiment of the tweet and the value given by majority of them are chosen as the sentiment of the feature. This method of feature selection limits the run time and memory usage in stream environment. Community detection and sentiment analysis were combined to study the sentiment of data in Online Social Networks (William and Hu 2013). These components integrated together, examines the structure and content of social network. Sanders corpus and Microsoft Corpus were used as data set for this study. By implementing this combination of methods, yields in an increase in modularity value.

A sentiment classifier is built to identify the polarity of the words namely positive, negative and neutral in microblogging applications (Saba et al., 2017). This methodology automatically extracts data from the corpus and identifies the polarity of the sentimental words irrespective of the size of the corpus. It builds a sentiment classification for microblogging by performing statistical linguistic analysis of the corpus. Tree Tagger is used for part of speech tagging and multinomial Naïve Bayes classifier is used in this approach which has been proved efficient compared to prevailing methods.

A new model for classifying tweets is proposed in six which combines prior polarity and tree kernel approach for feature engineering (Agarwal 2011). Sentiments are classified in two-way task and three-way task. The experiments are conducted in unigram, feature based and tree based methods. Feature based approach is deployed in this work which analysis approximately 100 features identified from tweets in microblogging sites and these features add marginal value to classifier. On the other hand emoticons are also taken into consideration and given equal preference as text. This approach has shown that prior polarity identifies the most important features from the blogs.

Comparative analysis on few existing approaches of opinion mining is discussed (Vishal and Sonawane 2016). Several machine learning and artificial intelligence algorithms like Naive Bayes, Max Entropy, and Support Vector Machine is applied to calculate the performance measure distinctly. Few methods in sentiment analysis are unveiled and the challenges were discussed (Karthikeyan and Vinothkumar 2016). A couple of metrics were evaluated on twitter dataset from Stanford University by applying numerous feature extraction techniques. Both the bigram and unigram approaches were used, and this study proves that accurate results can be obtained from cleaner data. It shows that there is a significant increase in the accuracy of performance metrics when bigram model is used compared to unigram model. The method handles tweets dynamically using a classifier based on the common features extracted from the tweets (Shenghua et 2005). Topic-Adaptive Sentiment Classification (TASC) is the semi supervised model which is framed to minimize the loss incurred in handling unlabeled data and features. The sentimental relationship between the tweets and emotion of the author is also captured using this topic adoptive feature. The topics as features are chosen from the tweets using collaborative method and TASC-t time line model is proposed to handle dynamic tweets. This method shows an impressive increase in performance metrics. A study to handle hash tags in twitter platform for sentiment classification is discussed (Wang et al., 2011). Hash tags are icon used as the prefix of keywords in tweets to signify the important terminology. It is represented by hash symbol. In this method, the polarity of the sentiment is timely automatically calculated relying on the hash tag. This differentiates this method from conventional sentiment analysis. It incorporates the synonym of the hash tag, the semantics of the sentence and relationship between the hash tags. A graph-based model is defined which automatically incorporates the tweets and hash tags and this gives better performance compared with base line approach.

Most of tweets in microblogging websites are ambiguous and sarcastic. A pattern-based approach is proposed to handle sarcastic comments (Mondher and Otsuki, 2016). This method recognizes the attitude and opinion of the user and discriminate the features into four categories by using classification techniques. The features that were collected include uncommon words, the count of the same, sarcastic expressions, count of interjections and laughing expressions. Around 6000 tweets were collected using the hash tag Sarcasm. By experiments it has obtained 83.1% for accuracy and 91.1% precision.

A study in academic and industrial data sets was done by Tan et al., (2014) to identify the foreground topics and the rank reason candidates using Latent Dirichlet Allocation (LDA) based model and Foreground and Background LDA (FB-LDA). The foreground and background topics were identified from the data sets and the study claims that foreground topic can give potential interpretations of sentiment variations. An elite set of representative tweets were studied using Reason Candidate and Background LDA (RCB-LDA) and the foreground topics were ranked based on popularity. This representative model identifies the intuition behind the disparity of opinion given by users.

The sentiment of the tweets was identified based on the occurrence patterns of words in their respective context using Senti Circle approach proposed by Hassan et al., (2015). The polarity strength, Senti-Strength of the word is updated in the sentiment lexicon. This method identifies the sentiment of the word in the entity level and tweet level too. Three huge tweet datasets were used for experimentation and this approach supersede the existing lexicon labeling methods in identifying the sentiment strength of the words. A lexicon-based classifier ensemble method identifies the polarity of tweets (silva et al., 2014, Hussein 2016). This ensemble combines Multinomial Naive Bayes, SVM, Random Forest, and Logistic Regression and thus improves the accuracy of classifier (Hirdoy et al., 2015). It has replaced the Bag of Words (BoW) routine by feature hashing and reduce the computational effort (Xing and Zhan 2015). Using this method, the users who count on the opinion shared in social network gain benefit in choosing their brands.

2. OUR METHODOLOGY

The dataset corpus used in our proposed method has around 25,000 reviews crawled from the web for a specific domain (Restaurant Chain).

2.1 REVIEW EXTRACTION FROM DOCUMENT

The subjective reviews are segregated for each commodity using data extraction. During the preprocessing phase we parse the reviews and remove the stop words. In addition the ambiguous words are identified not only based on their similarity but also based on semantic meaning. We have framed a new routine for preprocessing the reviews. The Fig 1 illustrates the design of our proposed approach. The corpus processes for linguistic analysis. Following which polarity mining and bigram analysis of data is done.

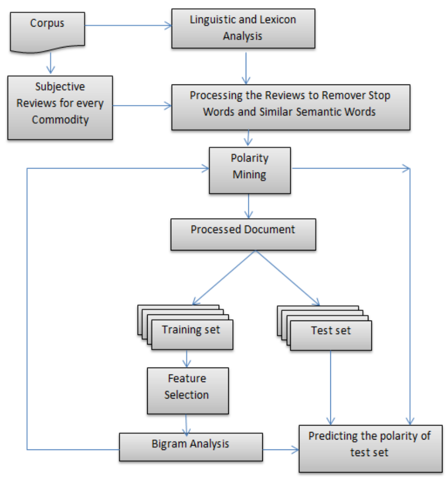

Figure 1. Design of our approach

During the preprocessing stage, the words or a string of words are identified using the bag of words model. We build a large list, a class of dictionary which encapsulates the words and sentiment. Each word has an associated sentimental value which is derived from crawling. We build a record of positive and negative words through web scraping and build the lexicon. The identified words and its sentiment are stored in the database as a single table with <word,

value> pair, where word refers to w^1, w^2, w^3 $w^4 \ldots w^n$ has an associated value series as v^1, v^2, v^3, $v^4 \ldots \ldots \ldots v^n$, n is the number of words extracted from the document. The value $v^i$of a word ranges from -1 to +1. In general, we consider words associated with negated values as negative words and positive values as positive words. This was implemented using Python and Sqlite tools. The crawled data set of reviews is stored as a file in database too. We evaluated the file against the dictionary of <word, value> pair and identified the polarity of the words in the file and also calculated the count of opinion words. The file of reviews is now replaced by a document containing the calculated information which is used for further analysis. The sentiment classifier in (Fig 2) recognizes the opinion word based on linguistic analysis.

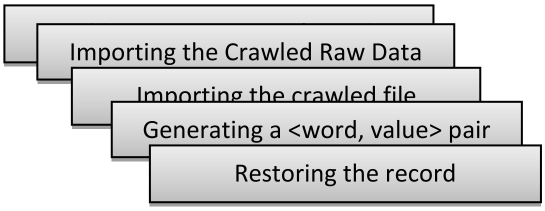

Figure 2. Building a sentiment classifier

2.2 Bigram

The crawled review set is accessed for bigram analysis. A sequence of two adjacent tokens or letters in a string of data is called a bigram. This frequency distribution of bigrams plays a major role in sentiment analysis to identify two consecutive positive or negative words in a single review. An example for bigram analysis is discussed below in Table 1.

Table 1. Bigram analysis example

Review Sentence: I like the beautiful red dress
Bigram input: $P(\text{I, like, the, beautiful, red, dress})$
Combinations: $P(\text{like}/\text{i}), P(the/\text{like}), P(beautiful/\text{the}), P(red/beautiful), P(dress/red)$

Table 2. A sample Bigram set

Bigram	**Sentiment**	**Original count**	**Ignored/Not Ignored**
food good	Positive sentiment	2400	Not ignored
like the	Positive sentiment	5000	Not ignored
thank lot	Positive sentiment	199	Not ignored
come tomorrow	Neutral	50	Not ignored
i got	Neutral	550	Ignored
here with	Neutral	390	Ignored
may be	Neutral	200	Not ignored
i have	Neutral	150	Not ignored
in today	Neutral	56	Ignored
by tomorrow	Neutral	29	Ignored
dressed red	Neutral	45	Ignored
Never go	Negative sentiment	60	Not Ignored
Waiter come	Neutral	16	Ignored

In the general way of bigram analysis all possible combination conditional probability values of $(i+1)^{th}$ word and $(i)^{th}$ word are identified. In our method we have already processed the document for linguistic analysis and removed all the unwanted information. The post-processing review sentence will be 'like beautiful red dress', where like is the opinion word which is taken to be positive based on method already discussed. So the sentence is refined as beautiful red dress. The bigram analysis when made on this sentence gives combinations as $P(like/i)$ $P(red/beautiful)$, $P(dress/red)$. This reduces the count of calculating the probability values by 12% leading to refinement of information. Table 2 lists the sentences encountered while processing the data set.

3. Training data set selection

The resultant data set contains the data pairs from bigram analysis and the opinion word's polarity. The features are chosen from the bigrams, as in bigram analysis all the unwanted pairs and adjectives are removed and is represented in Table 3. Thus the dataset is not crawled again to identify for the features. The resulting documents from the preprocessing stage comprise suitable information which specifically outfits the scope of our work. As procedures are carried out in the previous phase, we ensure that the dataset is beyond unwanted data leading to considerable increase in the performance.

Table 3. Count of Extracted Bigrams

Approach	**Pairs Count**
Regular Bigram Approach	9145
Excluding unwanted pairs of Bigram	8059

Among the 25,000 reviews $1/4^{th}$ (6250) of the features were used for training and $3/4^{th}$ (18750) were used for testing. In machine learning, the data sets are evaluated by dividing entire corpus into training and test set. A detailed study on the training set of the corpus yields better prediction in test set. To identify the idle training set, we divide the data set into four quarters. We calculate the performance measure; recall by considering every quarter as training set and the rest as test set. The obtained recall values through experiments are tabulated in Table-4. This procedure is done iteratively for every quarter. Recall identifies the goodness of a test which detects the positive tuples as positive. Following which precision, a measure identifies the quality of positively classified tuples. As a primary phase, we identify the recall value for every quarter chosen as training set. The outcome of the above process which returns high value of recall is set as training set and the data set is further evaluated. The Figure 3 shows the performance measures against the training sets chosen.

Table 4. Recall Values for Every Quarter

Quarter	True Positive	False Negative	False Positive	True Negative	Positive Recall
Training Set – 1st Quarter	5614	300	186	150	94.92
Training Set – 2nd Quarter	5111	493	250	396	91.20
Training Set – 3rd Quarter	4898	650	245	457	88.28
Training Set – 4th Quarter	5785	224	143	98	96.27

1 Feature selection

From the above Fig 3, quarter-4 is chosen for evaluating our system and with quarter-4 as training set we evaluate our model. The below table 5 shows the performance measures calculated for different feature count ranging from 2^9 to 2^{14}.

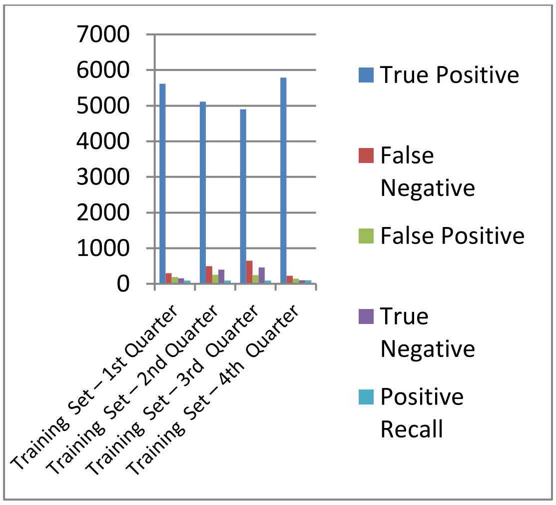

Figure 3. Performance measure against training set

It is inferred that as the number of chosen feature increases, the accuracy and the positive precision simultaneously increases for our system. It is to be taken into account, the negative precision value too as it plays major role in the quality of the system. When 4096(2^{12}) features were evaluated the accuracy value increased from 81.55 to 83.27, whereas the recall value increased from 0.322 to 0.413.

We consider the recall value as it decides the quality of the system. Along the iteration when 8192(2^{13}) features were evaluated the positive precision and accuracy showed a slight increase leading to decrease in negative precision. This count of features was considered idle for evaluating our system. The 8192 words as features contributed 2.18 percentage of the entire number of words in our data set (75000).

Table 5. Performance of the system against the feature count

Feature Count	Positive Precision	Positive Recall	Negative Precision	Negative Recall	Accuracy	F-Measure
29	0.845	0.834	0.264	0.067	73.7	84
210	0.891	0.88	0.326	0.061	80.56	88.6
211	0.905	0.879	0.322	0.061	81.55	89.2
212	0.893	0.912	0.413	0.063	83.27	90.31
213	0.918	0.917	0.382	0.052	85.52	91.79
214	0.912	0.916	0.419	0.06	85.14	91.5

Assuming an average of 15 words for every review expected that there are 75,000 words in our review data set. As along experiments the F-Measure value started increasing but it saturated at 2^{13} count of features.

4. Conclusion

In this work we have proposed a sentiment classifier which identifies the opinion of a word in minimal time duration in against the existing approaches. The reviews are processed to extract the bigrams. The root word and the word that qualifies the root word are taken into account. The model is evaluated with training set which is chosen based on the recall value. The system was tested with varying number of features in term of 2^n. We analyzed that the number of features extracted plays a major role as the system behaved effectively for a certain number of features. The accuracy of the system increased along with the number of features as well as the negative precision too. But after a certain count the negative precision value decreased giving respectable accuracy value. Experiments were performed for the above methods and the result description includes the values of performance measures.

5. REFERENCE

Agarwal A., B. Xie, I. Vovsha, O. Rambow and P. Rebecca, Sentiment Analysis of Twitter Data, Pro-ceedings of the Workshop on Language in Social Media Pp 30–38 (2011)

Hassan S., Yulan He, M. Fernandez, H. Alani, Contextual semantics for sentiment analysis of Twitter, Information Processing and Management Pp. 5–19 (2015)

Hridoy, S.A.A., M.T. Ekram, M.S. Islam, F. Ahmed and R.M. Rahman, Localized twitter opinion mining using sentiment analysis, decision analytics. A springer open Journal Pp. 1-19 (2015).

Hussein, D.M.E.D.M. A survey on sentiment analysis challenges, Journal of King Saud University – Engineering Sciences In Press (2016).

Karthikeyan, E.S. and R. Vinothkumar, Experimental study on stabilization of black cotton soil using textile sludge waste treated with hydrated lime. Pak. J. Biotechnol. 14(3): 411-415 (2017) .

Mondher B. and T. Otsuki, A pattern-based approach for sarcasm detection on twitter. IEEE Access Pp. 5477 – 5488 (2016).

Nathan A., J. Liddle and W. Hu, Twitter Sentiment in Data Streams with Perceptron, Journal of Computer and Communications Pp. 11-16 (2014)

Saba S.S., K.A. Sinaa, J.H. Reem, Attitude of a sample of general physicians working in some pri-mary health care centers in baghdad, al-karkh towa-rds family medicine, Pak. J. Biotechnol. 14(3): 429-436 (2017)

Shenghua L., X. Cheng, F. Li and F. Li, Tasc, Topic-Adaptive Sentiment Classification On Dyna-mic Tweets. IEEE Transactions On Knowledge And Data Engineering 27(6): 1696-1709 (2015).

Silva da, N.F.F., E.R. Hruschka and E.R. Hruschka Jr, Tweet sentiment analysis with classifier ensembles. Decision Support Systems Pp. 170–179 (2014).

Svetlana K., X. Zhu and S.M. Mohammad, Sentiment Analysis of Short Informal Texts. Journal of Artificial Intelligence Research Pp. 723-762 (2014).

Tan, S., Y. Li, H. Sun, Z. Guan, X. Yan and J. Bu, Interpreting The Public Sentiment Variations On Twitter, Ieee Transactions On Knowledge And Data Engineering 26(5): 1-14 (2014)

Vishal A.K. and S.S. Sonawane, Sentiment Analysis of Twitter Data: A Survey of Techniques, International Journal of Computer Applications 139 (11): 5-15 (2016).

Wang, X., F. Wei, X. Liu, M. Zhou and M. Zhang, Topic Sentiment Analysis in Twitter: A Gra-ph-based Hashtag Sentiment Classification Appro-ach, Proceedings of the 20th ACM international con-ference on Information and knowledge management Pp. 1031-1040 (2011).

William D. and W. Hu, Mutually Enhancing Community Detection and Sentiment Analysis on Twitter Networks. Journal of Data Analysis and Information Processing Pp. 19-29 (2013)

Xing F. and J. Zhan, Sentiment analysis using product review data. Journal of Big Data Pp. 1-14 (2015).