Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org
Volume 5, Special Issue 2, December (2015)

Sentiment Knowledge Discovery using Machine Learning Algorithms

Surbhi Bhatia
Research Scholar, Banasthali Vidyapith, Rajasthan, India

Manisha Sharma
Associate Professor, Banasthali Vidyapith, Jaipur, India

Komal Kumar Bhatia
Associate Professor, YMCA University of Science & Technology, Haryana, India

Abstract –With the increase of social networking, there has been a surge of user generated content online. Among all varieties of social media, Twitter is a valuable resource for data mining because of its popularity and acknowledgment by well-known persons. In this paper we present a system which collects Tweets from social networking sites and classify these tweets to provide some prediction of business intelligence. A flowchart has been proposed in which the overall picture of classification of twitter data has been proposed and accuracy of the evaluation strategies by various supervised learning algorithms has been evaluated. We propose a flowchart in which the overall picture of classification of twitter data has been illustrated from a popular real-time micro blogging service, Twitter, where users share their opinions. Results of trend analysis are evaluated in terms of basic information retrieval search strategies i.e. Precision and Recall.

Index Terms – Tweets, Sentiment Analysis, Machine Learning Algorithms, Web Mining.

This paper is presented at International Conference on Recent Trends in Computer and information Technology Research on 25th& 26th September (2015) conducted by B. S. Anangpuria Instituteof Technology & Management, Village-Alampur, Ballabgarh-Sohna Road,Faridabad.

1. INTRODUCTION

Now-a-days, various social networking sites like Twitter , Facebook , MySpace , YouTube have focused our research towards mining opinions and have become one of the most important applications of Web 2.0 [1]. Many people use Twitter as the media for sharing opinions on the web. Opinions can be defined as a private state of an individual represented in the form of emotions, sentiments, ideas etc [2].

Opinion Mining is the technique of detecting and extracting subjective information in text documents. Opinion mining refers to a sub discipline of computational linguistics that focuses on extracting people's opinion from the web [3].

Sentiment analysis on the other hand determines the contextual information, polarity (positive, negative or neutral) and polarity strength (weakly positive, mildly positive, strongly positive) of a document [4]. Opinion mining can be done at the Document, sentence and aspect (view) level. These tasks help to extract public opinion on feature of an entity. Classification is done based on four pairs of human emotions, i.e. joy-sadness, acceptance-disgust, Anticipant-Surprise, Fear-Anger [5]

Every day, about 250 millions messages are tweeted online. This diverts the focus of various companies who want to sell their products and analyze their status and brands. The organizations can analyze the sentiments of the tweets shared online in order to maintain their market status. The various machine learning algorithms can aid in extracting the sentiments and classifying them into positive and negative tweets.

The various machine learning algorithms used in our paper are discussed as follows:

1.1 Support Victor Machines (SVM): A Support Vector Machine (SVM) performs classification by finding the hyper plane that maximizes the margin between the two classes. The vectors (cases) that define the hyper plane are the support vectors [6].

1.2 Naïve Bayes Classifier (NB): Naive Bayes classifiers are a family of probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to

problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set [7].

1.3 Maximum Entropy (ME): The main idea behind maximum entropy principle is that unknown model generating the sample data should be the model that is most uniform and satisfy all constrains from sample data (or training data) [8].

The paper discusses the workflow of extracting sentiments online and getting to the final decision of accepting or rejecting a product. This is made possible by first extracting reviews from Twitter API and storing them in the repository.

Twitter messages posted as blogosphere are mostly expressed as informal text which require more processing as compared to formal text. Since, informal text consists of sarcasm, poor grammar, and non dictionary standard words [9], therefore preprocessing of the extracted tweets is required which includes removing non standard dictionary words, stop words, special characters etc. The features of the product are identified for classification using Senti word net dictionary. Thereafter various supervised learning algorithms are applied for getting these positive and negative reviews. Finally the evaluation is done for classification accuracy and results are shown, which is our main focus of the proposed work.

Our paper illustrated a complete framework and stresses on classifying entire documents according to the opinions on particular topic and then performance is measured by calculating accuracy.

The remainder of the paper is organized as follows. Section 2 presents the literature review. In section 3, work proposed is discussed. Section 4 presents the results and section 5 concludes.

2. RELATED WORK

Akshi Kumar [10] has determined the semantic orientation of the opinion words in tweets by using a hybrid approach using both corpus based and dictionary based methods with the help of a case study.

Anna Stavrianou [11] has presented a model based on opinion based graph which has focused towards content oriented domain. The proposed model has given a better technique of handling knowledge extracted from the discussion. The mining of the discussion has reduced the dimension space of the data.

Saif[12] has proposed the use of semantic features in Twitter sentiment classification . The paper has explored that the results on the semantic feature model outperforms the Unigram and POS baseline for identifying both negative and positive sentiment by conducting extensive experiments on twitter data sets.

Spencer[13] has proposed a tool called sentimentor for analyzing twitter data by using Naïve Bayes classifier. The comparison with the previous work is also achieved and analysis of the results confirmed that bigrams offer better performance with the classification process.

Hemalatha [14] presented a system which collects Tweets from social networking sites and analysis on those Tweets is being done. A result of trend analysis has been displayed as tweets with different sections presenting positive, negative and neutral.

Parikh [15] implemented two Naive Bayes unigram models, a Naive Bayes bigram model and a Maximum Entropy model to classify tweets.

Krzysztof [16] has discussed the new opinion classification method which is a variant of semantic orientation and presents the results by testing the algorithm for accuracy on data sets from real world.

Go[17] has discussed how to train sentiment classifier on Twitter data. The work has explored the use of several different classifiers across different n-grams with and without the use POS tags.

Mukherjee [18] has proposed a lightweight method for using discourse relations for polarity detection of tweets. The paper has shown that our discourse-based bag-of-words model performs well in a Twitter environment.

Pawel Sobkowicz [19] has proposed a new framework for opinion mining with respect to content analysis in social media. It has discussed the three important modules to track opinion data online. The further research has focused on the policy making issues through social media sources.

M S Vijaya [20] has discussed the importance and functionalities of different types of mining areas in social networks. It has emphasized on how opinion mining and sentiment analysis can be studied to gather knowledge which can promote business, political scenarios and other areas.

G.Vinodhini [21] has presented a systematic literature review of opinion mining techniques and methodologies and also explores the existing gaps in the field. The main challenging aspects in the paper exist in use of other languages, dealing with negation expressions; produce a summary of opinions based on product features etc.

Arti Buche [22] has focused on the achievement of the tasks of opinion mining. The problems faced in the sentiment

analysis for reviewing a product are discussed in order to provide a summarized framework of opinions.

Bing Liu [23] has explained the heuristic and rule based methods discussing the overall description of what opinion mining is, techniques used in sentiment classification and how opinion summarization can be performed.

3. PROPOSED WORK

The process of extracting and mining useful information or knowledge from web page content is called web content mining. Opinion mining is a part of web content mining where web pages can be clustered and classified automatically according to the different categories.

The generic framework of sentiment analysis is shown in Figure 1. The query posted by the user is preprocessed to the standard form in the form of DOM structure. The various social networking sites are searched for the web page containing opinions. Opinion downloader will perform parsing of the web page and downloads the opinions. The relevant opinions are searched from World Wide Web and stored in a buffer called as Repository.

Thereafter the opinions collected are passed over to another module. In the Opinion Classification module, the opinions are identified and classified by using Senti word dictionary for the sentiment analysis.

The positive or negative comments classified by various machine learning algorithms are evaluated and the accuracy of the same has been verified.

The overall architecture is discussed as follows:

3.1 User Query:

The user will post a query according to his her interest.

3.2 User Processing:

The query will be processed and various stop words removal, stemming, singular to plural etc will be performed to refine the query.

3.3 Web Database:

Web database will collect the data from various social networking sites, face book, twitter and other review sites.

3.4 Opinion Collector:

This module will download all the web pages from these specific sites in the HTML format and are stored in the Page Repository The links are also extracted which are stored further in the depth first manner.

3.5 Opinion Downloader:

This module will extract the relevant opinions from the stored web pages, ignoring the rest.

3.6 Opinions Repository:

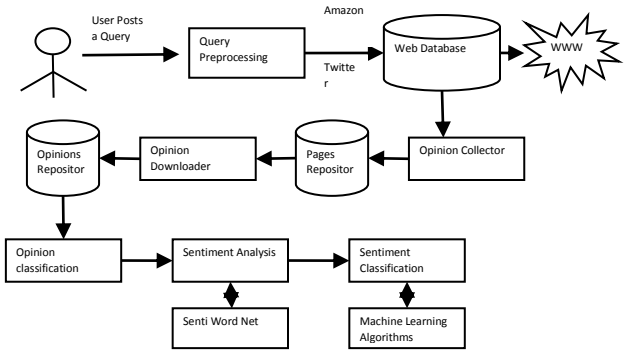
The result in the form of opinions is stored in this repository.

3.7 Opinion Classification:

To identify opinions, we first need to perform sentiment analysis. Senti word net is applied to calculate the score for each sentence, i.e. the positivity and negativity of the sentence is calculated. Sentiment Classification is done to classify opinions into positive and negative reviews at the document level. This is done by applying machine learning algorithms like Naïve Bayes Classifier, Support Victor Machines, and Maximum Entropy method.

Thereafter the classification done is evaluated with the performance metrics and the accuracy achieved is justified by taking data sets.

Figure1 Generic Framework



4. RESULTS AND DISCUSSIONS

The data sets generated using the data generation module. The experiments are conducted on the data set using twitter streaming API. Random tweets with dissimilar opinion words at different times were considered for analysis.

Our Data set consists of 810 tweets annotated by a group of 18 human annotators from which 410 have a positive polarity and 400 have a negative polarity. Our data set is collected by extracting the opinion word as Tablet. Precision, Recall and

F-Measure are used for evaluation of our proposed framework and comparison is done [24].

Precision is defined as the ratio of relevant tweets retrieved to the total number of tweets retrieved (relevant and irrelevant tweets retrieved. Mathematically,

Precision= RT/RT+RW

Where RT is the relevant tweets retrieved and RW is the irrelevant tweets retrieved.

Recall is defined as the ratio of relevant tweets retrieved to the manually retrieved tweets by the classifier (relevant tweets retrieved and relevant tweets not retrieved). Mathematically,

Recall= RT/RT+RN

Where RT is number of relevant tweets retrieved and RN are relevant tweets not retrieved.

F-Measure is the harmonic mean of both the precision and recall. Mathematically,

F-measure= 2* Precison* Recall/ Precision + Recall

We found that accuracy of Naïve Bayes classifier is much higher than the other two. It has been observed that Precision of crawling is high i.e. ranges from 83.56% to 92.9%, Recall of crawling process is also high i.e. ranges from 81.86% to 93.2% and F-measure of is also quite high i.e. from 82.70% to 93.04%.

5. CONCLUSION

One of the prime means of communication is micro blogging nowadays. In our research, we have presented an overall framework for sentiment analysis using twitters API and classified the tweets using various supervised learning algorithms and compared their performance. We have evaluated all the methods presented in this paper using Recall, Precision and F-Measure and found that the accuracy of Naïve Bayes classifier was much higher. The accuracy of 89.2% with 87.6% with precision and 86.3% with recall has been achieved. Our future work will include the development of web application by using unsupervised learning algorithms and supervised learning algorithms and comparing the performance of both the algorithms.

REFERENCES

[1] L. Colazzo, A. Molinari and N. Villa.2009. “Collaboration vs. Participation: the Role of Virtual Communities in a Web 2.0 world”, International Conference on Education Technology and Computer, pp.321-325.

[2] K. Khan , B. Baharudin , A. Khan, A. Ullah, 2014.”Mining opinion components from unstructured reviews: A review”,1319-1578.

[3] Bhatia, S., Sharma, M., Bhatia, K.., 2015. Strategies forMining Opinions: A SurveyInternational Conference on “Computing for Sustainable Global Development”, IEEE Xplore.

[4] Bahrainian, S.-A., Dengel, A., 2013. “Sentiment Analysis and Summarization of Twitter Data”, Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on DOI: 10.1109/CSE.2013.44, Page(s): 227 – 234, IEEE Xplore.

[5] Jędrzejewski, K., Morzy, M., 2011. “Opinion Mining and Social Networks : A Promising Match”, International Conference on Advances in Social Networks Analysis and Mining.

[6] Andrew, Ng., Part V. Support Vector Machines, cs229.stanford.edu/notes/cs229-notes3.

[7] Rish, Irina, " An empirical study of the naive Bayes classifier", 2001, IJCAI Workshop on Empirical Methods in AI.

[8] Cuong, Nguyen Viet, et al.2006. "A Maximum Entropy Model for Text Classification."The International Conference on Internet Information Retrieval.

[9] Bahrainian, S.-A., Dengel, A., 2013. “Sentiment Analysis and Summarization of Twitter Data”, Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on DOI: 10.1109/CSE.2013.44, Page(s): 227 – 234, IEEE Xplore.

[10] Akshi Kumar and Teeja Mary Sebastian. 2012.” Sentiment Analysis on Twitter”. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 4, No 3, July 2012 ISSN (Online): 1694-0814.

[11] Buche A., Chandak, M.B., Zadgaonkar, A., 2013. ” Opinion Mining and Analysis: A Survey”, International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June.

[12] Saif, Hassan, Yulan He, and Harith Alani. 2012. "Semantic sentiment analysis of twitter." The Semantic Web–ISWC, Springer Berlin Heidelberg, 508-524.

[13] Spencer, James, and Gulden Uchyigit. 2012. "Sentimentor: Sentiment analysis of twitter data." Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.

[14] I.Hemalatha , G. P Saradhi Varma, A.Govardhan, 2009. “Sentiment Analysis Tool using Machine Learning Algorithms”, International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), ISSN 2278-6856

[15] R. Parikh and M. Movassate, 2009.“Sentiment Analysis of UserGenerated Twitter Updates using Various Classification Techniques”, CS224N Final Report.

[16] Sobkowicz, P., Kaschesky, M., Bouchard, G., 2012. “Opinion mining in social media: Modelling, simulating, and forecasting political opinions in the web”, Volume 29, Issue 4, October, Pages 470 479.

[17] Go, A., Bhayani, R., Huang, L., 2009. "Twitter sentiment classification using distant supervision". Processing 150(12), 1–6, http://www.stanford.edu/~alecmgo/

[18] Mukherjee, Subhabrata, and Pushpak Bhattacharyya. 2012. "Sentiment Analysis in Twitter with Lightweight Discourse Analysis." COLING.

[19] Sobkowicz, P., Kaschesky, M., Bouchard, G., 2012. "Opinion mining in social media: Modelling, simulating, and forecasting political opinions in the web", Volume 29, Issue 4, October, Pages 470 479.

[20] Vijaya, M.S., Pream Sudha, V., 2013."Research Directions in Social Network Mining with Empirical Study on Opinion Mining", CSI Communications, December.

[21] Vinodhini, G., Chandrasekaran, R.M., 2012. "Sentiment Analysis and Opinion Mining: A Survey",Volume 2, Issue 6, June, ISSN: 2277 128 International Journal of Advanced Research in Computer Science and Software Engineering.

[22] Buche A., Chandak, M.B., Zadgaonkar, A., 2013. " Opinion Mining and Analysis: A Survey", International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June.

[23] Liu,B., 2012. "Sentiment Analysis and Opinion Mining", Morgan & Claypool Publishers May.

[24] Harris, Christopher G., 2012. "An Evaluation of Search Strategies for User-Generated Video Content." CrowdSearch.