ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 8, Issue, 10, pp.40433-40437, October, 2016

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

OPINION CLASSIFICATION SYSTEM USING SUPERVISED LEARNING ALGORITHM

***Swati Ambasta**

Dr. APJ Abdul Kalam Technical University, Lucknow

ARTICLE INFO

Article History:
Received 26th July, 2016
Received in revised form
25th August, 2016
Accepted 20th September, 2016
Published online 30th October, 2016

Key words:

Opinions Mining, Twitter,
Sentiment Analysis, Naive Bayes.

ABSTRACT

Due to huge amount of data posted online, decision making process considering opinions play a crucial role in everyone's life. Analysis in the field of making decisions and setting policies has shown that sentiment analysis and Opinion mining lies at the intersection of Question Answering system and Computational Linguistics. World Wide Web has tremendous amount of unstructured data present in web forums, social networking sites and other social platforms as reviews which diverts our study towards mining the opinions on web. Our research work focuses on extracting tweets, classifying them into positive, negative and neutral category and finally providing a recommendation as whether to buy or reject a product. Classification system is been proposed by using twitter data using Naive Bayes algorithm and accuracy of the evaluation strategies by has been evaluated. Review data is collected for various product domains from micro blogging sites like twitter, face book.

Citation: Swati Ambasta, 2016. "Opinion classification system using supervised learning algorithm", *International Journal of Current Research*, 8, (10), 40433-40437.

INTRODUCTION

World Wide Web has affected the way of making decisions on certain products because people posts online reviews immensely on Social media. Social Networking sites fascinate people to post feedbacks and reviews online on blogs, Internet forums, review portals and much more. These opinions play a very important role for customers and product manufacturers as they tend to give better knowledge of buying and selling by setting positive and negative comments on products and other information which can improve their decision making policies. Twitter messages posted online is about 250 millions per day which forces the organizations to observe their status and brands by extracting and analyzing the sentiments of the tweets shared online. To keep track of products and brands on the basis of their positive or negative views can be done through web using various supervised learning methods.Recent studies have been focused consistently and actively carried on to mine opinions. The primary objective of opinion mining is to categorically present these opinions to users illustrating the preference, featuring at the document level and then more precisely at the sentence level. The various machine learning algorithms used for classification are discussed as follows: 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 (cs229.stanford.edu/notes/cs229-notes3). Naïve Bayes Classifier (NB): A Support Vector Machine as stated by Luis *et al*. (Luis Gonz, 2005) (SVM) performs classification by constructing an N dimensional hyper plane that optimally separates the data into two categories (Ayodele, 2010). 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) (Cuong *et al*., 2006). Opinions can be defined as a private state of an individual represented in the form of emotions, sentiments, ideas etc. (Khan *et al*., 2014). Opinion mining refers to a sub discipline of computational linguistics that focuses on extracting people's opinion from the web (Bhatia *et al*., 2015). 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 (Osimo *et al*., 2008). 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 (Kamath *et al*., 2013)

The problem with the social media is that a large no of unstructured data is available. The aim is to present a framework where a summarized form of opinion reflects the decision making process for a user easier and efficient. The algorithm we develop is novel which is based on the extraction

****Corresponding author: Swati Ambasta,***
Dr. APJ Abdul Kalam Technical University, Lucknow

of features from the reviews. The tweets available on twitter as comments will be extracted which follows gathering of the relevant features from them and finally to categorize into three different categories, i.e positive, negative and neutral. The summarized opinion is presented based on the polarity scoring count. The classification is done using the supervised learning algorithm. We see that the algorithm proposed in our work out performs the previous work done. Informal text consists of sarcasm, poor grammar, and non dictionary standard words (Bahrainian and Dengel, 2013). After extracting tweets from Twitter API, preprocessing is done, the features of the product are identified and then Senti word net dictionary is used to assign polarity score count to all the extracted features. Thereafter supervised learning algorithm is 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.

Literature review

Opinion Mining is the technique of detecting and extracting subjective information in text documents (Bhatia *et al*., 2015). Kaiser *et al*. (2009) focuses on mining relationships in online communication. The relationships that exist between users are required to be analyzed in the formation of opinion online which is more specifically explained graphically with coherence in detecting opinion trends. The paper explains the major components which decide over opinion trends i.e. density and randic connectivity. It has the limitation that the proposed algorithm will fail to give correct opinion trends in dynamic social network analysis. Also contextual dependency and vague and complex sentences need further attention. Diana Maynard (2012) focuses on entity centric approach and makes use of linguistic relations on a rule based system. It also discusses about the various tools used in opinion mining analysis and the related issues. The method presented in the work can be easily adapted to the new tasks, domains and languages. The linguistic components discussed can be taken for preprocessing initially for machine learning tasks. With the advantages discussed, the paper has certain disadvantages as well. First is that the author has not evaluated any of the opinion mining work with tools. Second it does not work well on abbreviated text like those given in tweets. Third resolving ambiguity in formal text statements is still a problem. Farhan Hassan Khan (2014) presents a novel three way classification algorithm for twitter analysis. The results are taken by performing experiments using random tweets collected which are proved mathematically more accurate removing the limitations of sarcasm and sparsity. The proposed comprehensive TOM framework overcomes the limitations in the previous sentiment analysis papers. An important feature of the proposed architecture is that the raw data is efficiently processed and mistakes, abbreviation and noise is removed before providing input to the classifier. Preprocessing of data is still a problem with respect to time. Isidro Peñalver-Martinez (2014) presents a new approach of feature based opinion mining in semantic web technologies. It combines the use of domain ontology with domain independent sentiment analysis processes. The approach presented does not require any human intervention as it is fully automatic, manipulation of ontology allows feature selection to be applied on any domain which makes it domain independent and language independent. The drawback is ontology construction which is difficult and consumes a lot of time. Validation should be extended involving other domain which should be made dynamic. Pawel Sobkowicz (2012) 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. Vijaya (2013) 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. It has further promised to deploy the techniques developed by the research community in real world applications. Vinodhini (2012) has presented a systematic literature review of opinion mining techniques and methodologies and also explores the existing gaps in the field. The future work has included issues to resolve performance measures in sentiment analysis. 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 (2013) 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 (2012) 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. The limitation of the work is that more work needs to be done on probabilistic methods.

Proposed work

Web content mining refers to the process of extracting and mining useful information or knowledge from web page contents. Opinion mining comes under the category of web content mining wherein we can automatically classify and cluster web pages according to the topics. The generic framework of sentiment analysis is shown in Fig 1. First, the user posts a query according to his her interest. Then, the query will be preprocessed in order to remove the unwanted content. The query is searched within various social platforms and WWW will be searched for the particular HTML Page and parsing is performed, where Opinion Extractor will extract the relevant opinions and store them in a buffer called as Result Opinions Repository. Thereafter the opinions collected are passed over to another module. In the Opinion Identification 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:

User Query

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

User Processing

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

Web Database

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

Opinion Retriever

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.

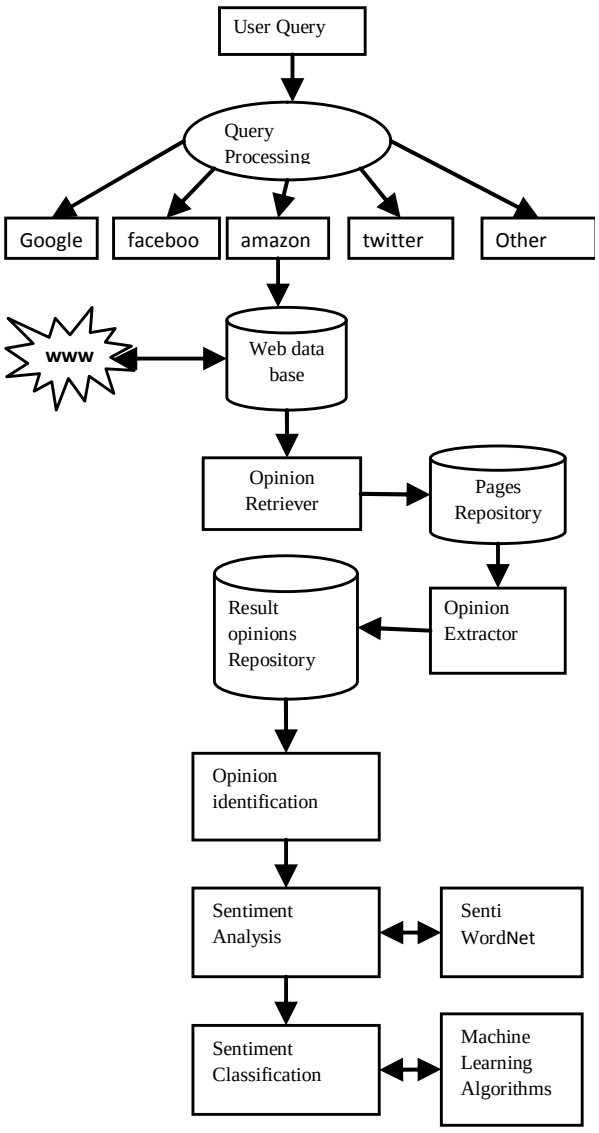

Fig. 1. GenericFramework

Opinion Extractor

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

The working of the opinion extractor is given in Figure 2.

Result Opinions Repository

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

Opinion Identification

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 algorithm like Naïve Bayes Classifier. Thereafter the classification done is evaluated with the performance metrics and the accuracy achieved is justified by taking data sets. Since we are only focusing on English tweets, so we will ignore rest of the tweets in other languages. The working of our system is represented in a flowchart given in Fig 2.

RESULTS

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 670 tweets annotated by a group of 21 human annotators from which 460 have a positive polarity and 210 have a negative polarity. Our data set is collected by extracting the opinion word as Samsungphone. Precision, Recall and F-Measure are used for evaluation of our proposed framework and comparison is done. Precision is defined as the ratio of relevant tweets retrieved to the total number of tweets retrieved (relevant and irrelevant tweets retrieved. Mathematically,

Precision= RTT/RTT+RWT

Where RTT is the relevant tweets retrieved and RWT 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= RTT/RTT+RNT

Where RTT is number of relevant tweets retrieved and RNT are relevant tweets not retrieved.

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

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

The evaluation results of precision, recall and F-Measure of complete dataset of Naïve Bayes classifier is shown in Table 1.

Table 1. Evaluation Results

No. of tweets	Precision (%)	Recall (%)	F-Measure (%)
Data Set 1	83.56%	81.86	82.70
Data Set 2	85.7	85.86	85.77
Data Set 3	87.82	89.9	88.84
Data Set 4	91.27	93.73	92.48
Data Set 5	92.9	93.2	93.04

We found that accuracy of Naïve Bayes classifier is much higher than the other supervised learning methods like SVM, Maximum Entropy etc. The results shown above verify the superiority of the above classifier. It has been observed that everything we see online, but we want to maximize the extraction of the data in a timely manner, so that the recent tweets posted gets updated frequently. This can be done by using web crawler. Our future work will be based on this

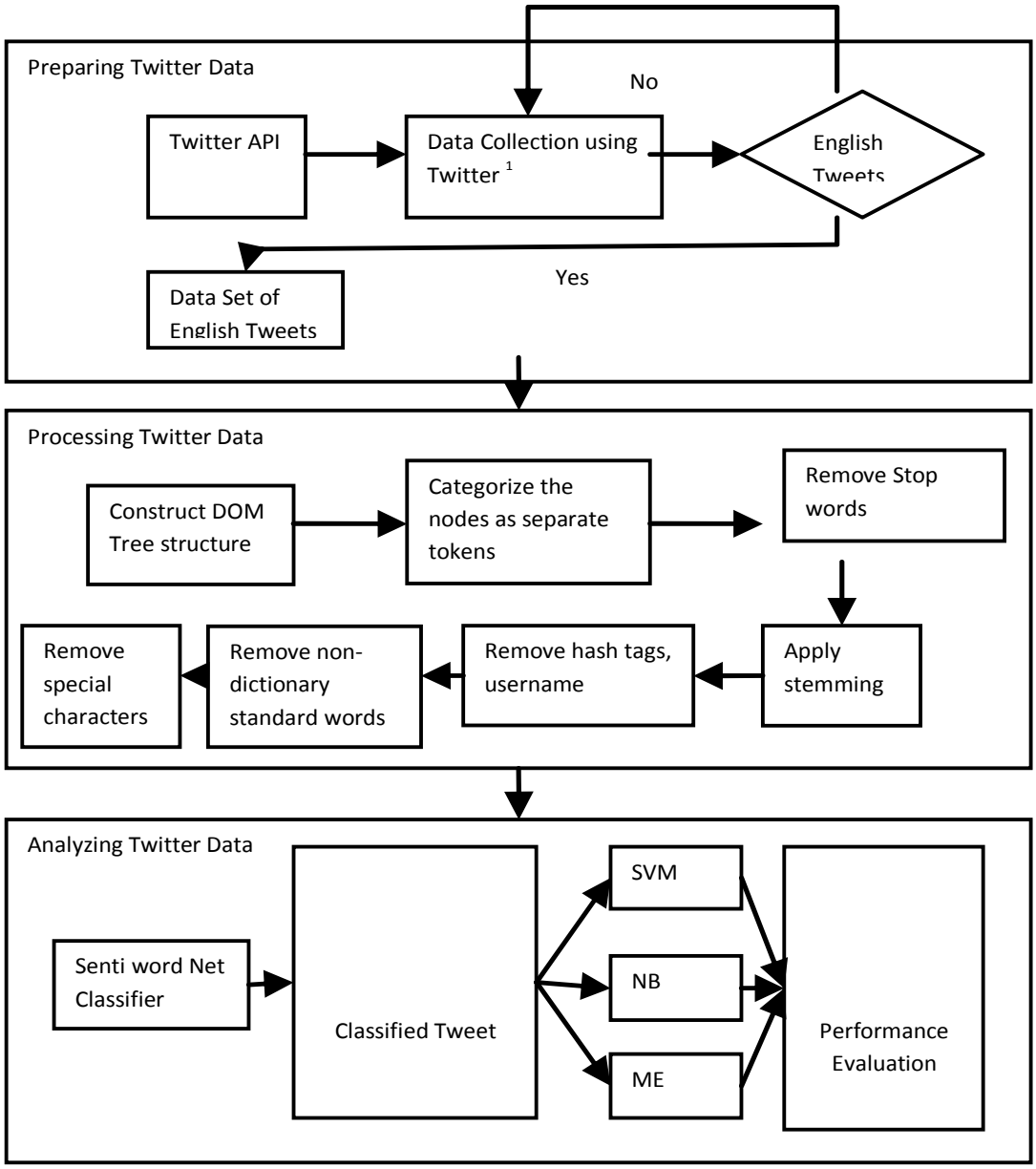

Fig. 2. Flowchart of our system

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% to93.2% and F-measure of is also quite high i.e. from 82.70% to 93.04%.

Conclusion

Micro blogging is the prime means of communication in today's world. Our system provides new features and benefits in recommendation of a product by considering reviews. This paper presents a research plan with an overarching goal to help ensure that the proposed opinion system is developed. 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 than the other two supervised learning algorithms i.e. Support Vector Machines and Maximum Entropy. The main problem is that we can't fit development. Another limitation is that the data extracted needs to preprocessed before classifying into positive, negative and neutral category. Following the line of this work, we will carry on to work further on the opinion ranking and summarization methods.

REFERENCES

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

Ayodele, T.O. Types of Machine Learning Algorithms, Feb 1, 2010.

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.

Bhatia, S., Sharma, M., Bhatia, K., Strategies forMining Opinions: A SurveyInternational Conference on "Computing for Sustainable Global Development", IEEE Xplore, 2015

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.

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

Kaiser, C., F. Bodendorf ,"Opinion and Relationship Mining in Online Forums",Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences Volume: DOI: 10.1109/WI-IAT.2009.25, Page(s): 128 – 131, IEEE Xplore, 2009.

Kamath, Bagalkotkar, S.S.,Kandelwal, A., Pandey, A., Poornima, S., 2013. "Sentiment Analysis Based Approaches for Understanding User Context in Web Content", CommunicationSystemsand Network Technologi es (CSNT), International Conference on DOI: 10.1109/ CSNT.130 , Page(s): 607 – 611, IEEE Xplore.

Khan, F.H., S. Bashir, U. Qamar,"TOM: Twitter opinion mining framework using hybrid classification scheme", Decision Support Systems, Volume 57, Pages 245-257, Science Direct.January 2014.

Khan, K., B. Baharudin, A. Khan, A. Ullah,"Mining opinion components from unstructured reviews: A review", 1319-1578 2014, 2012.

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

Maynard, D., K. Bontcheva, D. Rout,"Challenge in developing opinion mining tools for social media", workshop at LREC, 2012.

Osimo, D., Francesco, M., Anderson, C. 2008. "Research Challenge on Opinion Mining and Sentiment Analysis", Wired Magazine, 16(7), 16–07.

Peñalver-Martinez, I., F. Garcia-Sanchez, R. Valencia-Garcia, M. Á. Rodríguez-García, V. Moreno, A. Fraga, Jose Luis Sánchez-Cervantes", Feature-based opinion mining through ontologies",Expert Systems with Applications, Volume 41, Issue 13, 1, Pages 5995-6008, Science Direct, October 2014.

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 470479.

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

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.

****