International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 10, 2020
ISSN: 1475-7192

High Dimensional Data Sets Using Advanced Data Engineering Techniques for Sentiment Analysis

[1]ShashidharaniVaddineni, [2]HanyEldeib

Abstract

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service. The difficulties of performing sentiment analysis in this domain can be overcome by leveraging on common-sense knowledge bases. Opinion Mining is a very challenging and promising discipline which is defined as an intersection of information retrieval and computational linguistic techniques to deal with the opinions expressed in a document. The main aim at solving the problems related to opinions about products, reviews ranking in movies, Politian in newsgroup posts, review sites etc. In this paper we are about to cover the source of data from where we take , its classification, evaluation process and then grouping techniques, tools used, and future challenges in opinion mining. Opinion mining consists of various stages such as extraction of data from various sources, text classification, grouping together and then evaluating it to positive or negative or true or false value. On the basis of our survey and analysis of the techniques, we provide an overall picture of what is involved in developing a software system for opinion mining. Any user, buyer or customer rely on the Web for their opinions on various products and services which they have used, it is very important to develop methods to automatically classify and evaluate them. The task of classifying and analyzing such collective data together is known as customer feedback or review data, and is called as opinion mining.

Keywords: *Opinion Mining, Data Engineering, Sentiment analysis*

I. Introduction:

The World Wide Web is growing at an alarming rate not only in size but also in the types of services and contents provided. Each and every users are participating more actively and are generating vast amount of new data. These new Web contents include customer reviews and blogs that express opinions on products and services –

[1] *Virginia International University*
[2] *Virginia International University*

which are collectively referred to as customer feedback data on the Web. As customer feedback on the Web influences other customer's decisions, these feedbacks have become an important source of information for businesses to take into account when developing marketing and product development plans. This era is of automated systems [1] and digital information every field of life is evolving rapidly and generating data. As a result huge amount of data produce in the field of science, engineering, medical, marketing, finance etc [2]. Automated systems are needed to automate analysis, summarization, and classification of data. It also helps at enterprise level to take related decisions. Multiple research fields like statistics, machine learning, artificial intelligence and visualization are involved to develop such automated systems

Opinion mining is a growing field to identify the thoughts and sentiments of people, which they express in form of their feedbacks or reviews on various things. Today due to vast use internet and social platforms, people are having a huge amount of space where they can publically express their opinions. These reviews are present in various forms on web like the feedbacks for products listed on various ecommerce web sites, or the personal posts from Face book, twitter, bloggers etc. Some formal reviews are also available in various discussion forums related to products/sites or domains. People also post a lot of personal views in form of movie reviews or the buzz creating news in various articles for magazines and newspapers. These opinions are directly related to how they feel. And this feeling can be classified as being positive, negative or neutral in nature. Positive views have a positive impact on society and a negative view creates a negative impact as shown in figure 1.

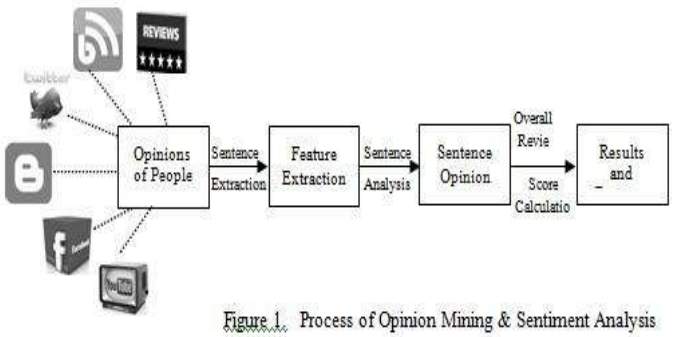
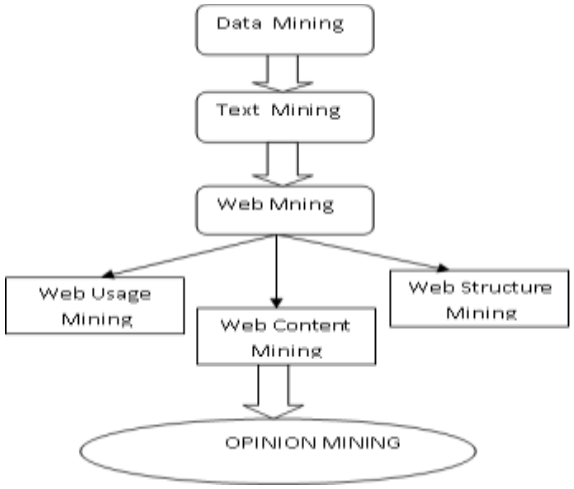


Figure 1. Process of Opinion Mining & Sentiment Analysis

Sentiment analysis of online user generated content is important for many social media analytics tasks. A lot of work has been carried out for extracting people sentiments from textual data. Researchers have largely relied on textual sentiment analysis to develop systems to predict political elections, measure economic indicators, and so on. Although, social media is source of most recent information, it cannot be trustworthy as it is composed of several aspects generated by different peoples. In this work we are proposing hybrid approach of sentiment analysis for area of interest. The hybrid approach consists of aggregating sentiments from both social media and news feeds. After extracting sentiments from both approaches, they are then clustered and will be made available for analysis. RSS feeds enable publishers to syndicate data automatically. A standard XML file format ensures compatibility with many different machines/programs. RSS feeds also benefit users who want to receive timely updates from favorites websites or to aggregate data from many sites.

A number of proficient ways are existing [4] to store the huge volumes of data, computational techniques and models are required to extract the hidden patterns and knowledge. These techniques and tools are used to transform the data into useful information, to make market analysis, fraud detection and find the customer intentions etc. Such computational tools and techniques are the subject of *Knowledge Discovery in Database andData Mining* [4-5].Text mining is an interdisciplinary method used in different fields like machine learning, information retrieval, statistics, computational linguistic and data mining to form mining algorithms. Some researchers defined text mining as tool to discover the new knowledge from huge volume of natural language text using computational algorithms. Web mining is a sub discipline of text mining used to mine the semi structured web data in form of web content mining, web usage mining and wed structure mining.



II. METHODOLOGIES

There are various methods used for opinion mining and sentiment analysis among which following are the important ones:

(i) Naïve Bays Classifier.

(ii) Support Vector Machine (SVM).

(iii) Multilayer Perceptron.

(iv) Clustering.

Categorization of work done for feature extraction and classification in opinion mining and sentiment analysis is done. In addition to this, performance analysis, advantages and disadvantages of different techniques are appraised Advantages & Disadvantages of above system are as follows. Advantages of Naïve Bayes Classification Method are Model is easy to interpret and Efficient computation. Disadvantage of Naïve Bayes Classification Method is Assumptions of attributes being independent, which may not be necessarily valid.Advantages of Support

Vector Machine Method are very good performance on experimental results and Low dependency on data set dimensionality. Disadvantages of Support Vector Machine Method are One disadvantages of SVM is i.e. in case of categorical or missing value it needs pre-processed and difficult interpretation of resulting model.

III. PROPOSED SYSTEM.

Sentiment analysis of online user generated content is important for many social media analytics tasks. A lot of work has been carried out for extracting people sentiments from textual data. Researchers have largely relied on textual sentiment analysis to develop systems to predict political elections, measure economic indicators, and so on. Although, social media is source of most recent information, it cannot be trustworthy as it is composed of several aspects generated by different peoples. In this work we are proposing hybrid approach of sentiment analysis for area of interest. The hybrid approach consists of aggregating sentiments from both social media and news feeds. After extracting sentiments from both approaches, they are then clustered and will be made available for analysis.

In this method we can take real time RSS feed data along with twitter data and then analyze that data to get the opinion. The main scope of this will be Grab the real time news data stream from Twitter using twitter streaming API 2. Grab the real time news from news RSS feeds

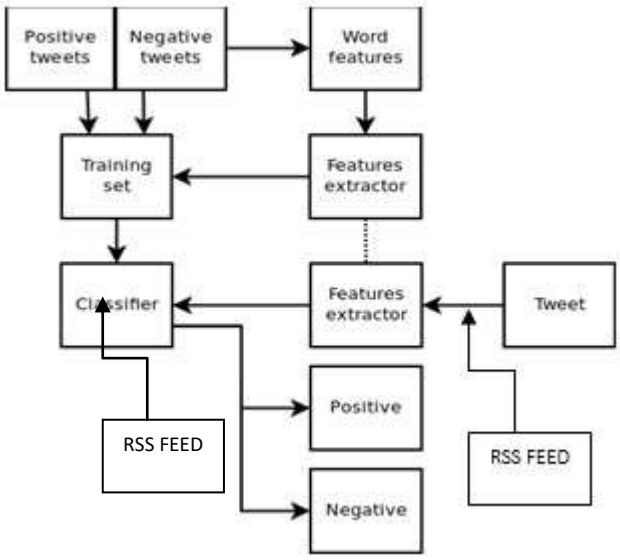


Figure 4.1.Proposed Working Block Diagram

As shown in figure 4.1 the feature will be extracted from RSS feed data and the Tweeter data which will be categories in different ways as Sport, Political etc. Then we can apply this with training data set and then it can be classified and the correct opinion can be provided.

This paper is organized as follows: section 2 covers opinion mining. Section 3 is about the dataset source. Section 4 is about the levels of sentiment classification. Section 5 describes text classification Section 6 is all about the grouping feature Section 7 describes evaluation process and Section 8 describes various recent tools used to do this

IV. Opinion Mining (O.M)

Opinion Mining is a promising discipline which is defined as combination of information retrieval and computational linguistic techniques deals with the opinions expressed in a document. The field major goals is solving the problems related to opinions about products, politics in newsgroup posts, review sites, etc. There are different techniques for summarizing customer reviews like Data Mining, Information Retrieval, Text Classification and Text Summarization [2], before World Wide Web users asked the opinions of his family and friends to purchase the product . In the very same way when any organizations need to take the decision about their products they had to conduct various surveys to the focused groups or they had to hire the external consultants to do so [4]. Web 2.0 [7], ease the customers to take decision to purchase the product by reviewing the posted comments. Customers can post reviews on web communities, discussion forums, twitters, blogs, product's web site these comments are called user generated contents. Web2.0 is playing a vital role in data extracting source in opinion mining. It facilitates users to know about the product from other customer's reviews who have already used it instead of asking friends and families. Companies, instead of conducting surveys and hiring the external consultants to know about the clients opinions, extract opinionated text from product web site [8]. An automated opinion summarization model is needed to complete these tasks. Opinion Mining or SentimentAnalysis is the area to extract the opinionated textdatasets and summarize in understandable form for end user [8]. Opinion mining is used to extract the positive, negative or neutral opinion summary from unstructured data. It involves subjectivity in text and computational management of opinion. It is the sub-discipline of web content mining, which involves Natural Language Processing and opinion extraction task to find out the polarity of any product consumers feedback [4]. Figure 2 describes the object model of Opinion Mining.

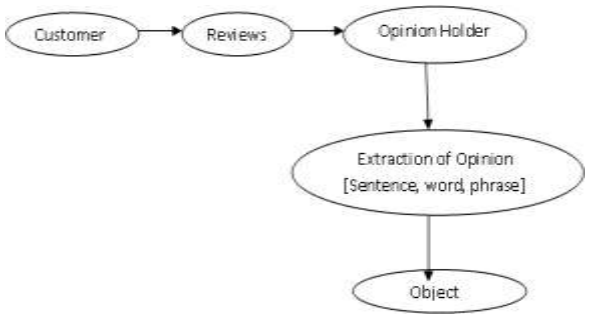


Fig. 2. Opinion Mining Model

Sentiment Classification

4.1 Document level:

Document level sentiment classification is based on the sentiments executed on the overall sentiments expressed by authors. Documents classified according to the sentiments instead of topic. It is very useful in summarizing the whole document as positive or negative polarity about any object (camera, fridge, mobile, car, movie, and politician). In [12] authors proposed a new approach "classification of opinion documents by a vote system" based on combining text representations using key-words related to bigrams. Sentiment Classification Using Phrase Patterns in used Special tags opinion words. System constructed some phrase patterns and compute sentiment orientation using unsupervised learning algorithm. Proposed system achieved 86% accuracy. Investigated perspective from which a document was written. They build Naïve Bayes based model and test on Israeli-Palestinian conflict. Their corpus consists of articles published on the bitter lemons website. They used NB-B (full Bayesian inference) and NB-M (Maximum a posteriori).

4.2 Sentence level:

Sentence level sentiment classification models is used for the extraction of the sentences contained in the opinionated terms, opinion holder and opinionated object. It is one level deep to document level and just concerns to the opinionated words but not the features. Total number of positive and negative words are counted from the extracted and classified sentences and if positive words are maximum then opinion about object is positive and if the negative words are more than opinion object is negative otherwise the opinion object will be neutral. To mine the customer reviews on a product proposed unsupervised algorithm is used and in this the algorithm find frequent features using Apriori algorithm. Chinese WordNet set classify opinion words in clauses (pos, neg or neutral) to summarize the comments. Sentence level opinion mining uses subjective and polarity (orientation) to find strength of opinions at the clause level. [13], all these are a notable work in this regard. To find the strength of opinions a new idea of syntactic clues is used. They use a wide range of features to find the strength of opinions.

V. CONCLUSION

The important part of gathering information always seems as, what the people think. The rising accessibility of opinion rich resources such as online analysis websites and blogs means that, one can simply search and recognize the opinions of others. One can precise his/her ideas and opinions concerning goods and facilities. These views and thoughts are subjective figures which signify opinions, sentiments, emotional state or evaluation of someone. In our proposed system we are using RSS feed data with real time twitter data base on the category.

VI. ACKNOWLEGMENT

I would like to show my gratitude to the Prof. Sonwane V.R project guide for sharing their pearls of wisdom with me during the course of this research. I am also immensely grateful to Prof.Shaikh I.R for their

comments on an earlier version of the manuscript, although any errors are my own and should not tarnish the reputations of these esteemed persons

REFERENCES

[1] Xue Li, , Vasu D. Chakravarthy, , Bin Wang, and Zhiqiang Wu, "Spreading Code Design of Adaptive Non-Contiguous SOFDM for Dynamic Spectrum Access" in IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, VOL. 5, NO. 1, FEBRUARY 2011

[2] J. D. Poston and W. D. Horne, "Discontiguous OFDM considerations for dynamic spectrum access in idel TV channels," in Proc. IEEE DySPAN, 2005.

[3] R. Rajbanshi, Q. Chen, A.Wyglinski, G. Minden, and J. Evans, "Quantitative comparison of agile modulation technique for cognitive radio tranceivers," in Proc. IEEE CCNC, Jan. 2007, pp. 1144–1148.

[4] V. Chakravarthy, X. Li, Z. Wu, M. Temple, and F. Garber, "Novel overlay/underlay cognitive radio waveforms using SD-SMSE framework to enhance spectrum efficiency—Part I," IEEE Trans. Commun., vol. 57, no. 12, pp. 3794–3804, Dec. 2009.

[5] V. Chakravarthy, Z. Wu, A. Shaw, M. Temple, R. Kannan, and F. Garber, "A general overlay/underlay analytic expression for cognitive radio waveforms," in Proc. Int. Waveform Diversity Design Conf., 2007.

[6] V. Chakravarthy, Z. Wu, M. Temple, F. Garber, and X. Li, "Cognitive radio centric overlay-underlay waveform," in Proc. 3rd IEEE Symp. New Frontiers Dynamic Spectrum Access Netw., 2008, pp. 1–10.

[7] X. Li, R. Zhou, V. Chakravarthy, and Z. Wu, "Intercarrier interference immune single carrier OFDM via magnitude shift keying modulation," in Proc. IEEE Global Telecomm. Conf. GLOBECOM , Dec. 2009, pp. 1–6.

[8] Parsaee, G.; Yarali, A., "OFDMA for the 4th generation cellular networks" in Proc. IEEE Electrical and Computer Engineering, Vol.4, pp. 2325 - 2330, May 2004.

[9] 3GPP R1-050971,"R1-050971 Single Carrier Uplink Options for EUTRA: IFDMA/DFT-SOFDM Discussion and Initial Performance Results ",http://www.3GPP.org,Aug 2005

[10] IEEE P802.16e/D12,'Draft IEEE Standard for Local and metropolitan area networks-- Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems', October 2005

[11] 3GPP RP-040461, Study Item: Evolved UTRA and UTRAN, December 200

[12] R. Mirghani, and M. Ghavami, "Comparison between Wavelet-based and Fourier-based Multicarrier UWB Systems", IET Communications, Vol. 2, Issue 2, pp. 353-358, 2008.

[13] R. Dilmirghani, M. Ghavami, "Wavelet Vs Fourier Based UWB Systems", 18th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, pp.1-5, Sep. 2007.

[14] M. Weeks, Digital Signal Processing Using Matlab and Wavelets, Infinity Science Press LLC, 2007.

[15] S. R. Baig, F. U. Rehman, and M. J. Mughal, "Performance Comparison of DFT, Discrete Wavelet Packet and Wavelet Transforms in an OFDM Transceiver for Multipath Fading Channel,", 9th IEEE International Multitopic Conference, pp. 1-6, Dec. 2005.

[16] N. Ahmed, Joint Detection Strategies for Orthogonal Frequency Division Multiplexing, Dissertation for Master of Science, Rice University, Houston, Texas. pp. 1-51, Apr. 2000.