Data Mining and Machine Learning: Implementation model, Applications, Challenges and Issue in Sentiment Analysis System - A Survey

[1] Mr.Ashish Kumar Soni, [2] Ms.Shweta Shah, [3],Ms.Harsha Solanki
[1]Assistant Professor, [2]Assistant Professor, [3]M.Tech Student
[1]Computer Science and Engineering Dept.,
Malwa Institute of Science and Technology,Indore

Abstract:

Sentiment analysis is the process of determining the opinion or filling of a piece of text. Companies across the word implementing machine learning to do it automatically. It is a supper use of opinion inside the text. Once we understand how the customer fills after analyzing their comments and reviews. One can identify what kind of things they like or dislike and built things like recommendation system. Or more target marketing systems for them. This paper discussed the different implantation model of sentiment analysis system with their advantage and disadvantages. This paper also included the some real world application where sentiment analysis system can be used. Different implementation challenges and other issues related to sentiment analysis system explored in this paper.

Keywords: Sentiment analysis, opinion mining, bag of words, NLP, AI, text mining, lexicon, machine learning, twitter.

1. Introduction:

Human emotions intelligence distinguishes them from every other non living being on the earth. These emotions can be simple like any one can get so angry. Human invented language to help them to express emotions to others [2]. But some time words are not enough. Some emotions do not have direct communication language translation. Emotions are heard to express. But that's where artificial intelligence (AI) and machine learning (ML) can help. [3] AI can help to understand emotions perhaps better than human being by analyzing emotional data to help in taking optimum decision. This can work as a personal life coach.

Rest of the paper organization is as follows. In section 2 different types of sentiment analysis systems are discussed with their comparative study. In section 3 two main approaches of sentiment analysis is compared. In section 4 elaborated the use of tweeter for sentimental analysis. Section 5 discussed some real word application in which sentiment analysis can be used. In section 6, this paper explored some challenges while development of any sentiment analysis system. Section 7 concluded the overall study.

2. Sentiment Analysis Systems:

There are generally two approaches to do sentiment analysis.

2.1 Lexicon Based Sentiment Analysis:

[4]The first one is the lexicon based approach. In this approach given text are divided into smaller words, phrases and sentences called token and this process is known as tokenization. [5]Then number of words and their frequency is counted. This resulting tally is called Bag of Word model. Next processes look up the subjectivity of each word from an existing lexicon, which is a database of emotional values of words. These words are pre recorded by researchers by this values analyzer can compute the overall subjectivity on text.[6] There are mainly three different approaches identified for lexicon based sentiment analysis. First is dictionary based approach [7] in this a small set of opinion words is collected manually with known orientations. The newly found words are added to the list then the next iteration starts. This iterative process continued till new word found. Manual checking is required to check the errors in this process. There is another approach, Corpus-based approach. This approach

analyzes sentiments by finding opinion words with context specific orientations. [8]This method depends on syntactic patterns or patterns that occur together along with a list of opinion words to find other opinion words in a big corpus [9]. Some time lexicon based natural language processing techniques are used for sentiment analysis [10].

2.2 Machine Learning Based Sentiment Analysis:

[11]The other approach uses machine learning. If there are some text labeled with positive or negative. We can train a classifier on it and give a new text to classifier to take decision weather text are positive or negative. This approach works on semantic of word. Machine learning based sentiment analyzer is difficult to implement with compare to lexicon based sentient analysis system. [12]Machine learning based sentiment analysis can be further classified in supervised and unsupervised learning. Supervised learning based on existing labeled documents. These documents give supervision platform to upcoming text. There are many classifiers which may be used in supervised learning like Probabilistic classifiers (Naı¨ve Bayes Classifier , Bayesian Network , and Maximum Entropy Classifier), linear classifiers (Support Vector Machines Classifiers and Neural Network), Decision tree classifiers and Rule-based classifiers [13]. The main purpose of text classification is to classify documents into a certain number of predefined categories. In order to accomplish that, large number of labeled training documents are used for supervised learning, as illustrated before. In text classification, it is sometimes difficult to create these labeled training documents, but it is easy to collect the unlabeled documents. The unsupervised learning methods overcome these difficulties. Many research works were presented in this field including the work presented by Ko and Seo [2]. They proposed a method that divides the documents into sentences, and categorized each sentence using keyword lists of each category and sentence similarity measure.

Other than these two approaches (Lexicon based and Machine Learning) some researchers also proposed few approaches for development sentiment analysis system. Wille [11] proposed a sentiment analysis technique called formal concept analysis. This technique uses mathematical approach for structuring, analyzing and visualizing data. One another approach called Fuzzy formal concept analysis was proposed for the unformatted and unclear information [12]. Mudinas et al. [13] proposed another concept level approach. This approach is a concept- level sentiment analysis system that is integrated with opinion mining lexicon-based learning approach. Cambria and Havasi [14] proposed a publically available semantic and affective resource for opinion mining and sentiment analysis. They develop a system called senticNet 2.

3. Lexicon based Vs Machine Learning

[15]As discussed in previous section there are two main approaches for sentiment analysis one is lexicon based approach and another next one is machine learning based. This section discussed the comparative study of both approaches.

Using a lexicon based algorithm is easy but machine learning approach is more accurate. There are several things in language that means show something but really means another. But the deep neural network understands the several things because they don't analyze only face value of text. They create abstract representation of what they created. This generalization is called vectors. Machine learning uses them to classify data.

4. Sentiment Analysis through Twitter

[14]Twitter is a treasure tour of sentiments. People around the word put thousands of reaction and opinion on every topic under the sun every day. It's like one big psychological database and continuously being updated. We can be used it to analyze millions of text in seconds with the power of machine learning.

Sentiment analyzer receives some input text like twitter twits. Firstly the text has to split into several words or sentences. This process is called tokenization, because this process creates small tokens form big text. The process just count the each words shows up once the text is tokenized. This is called bag of words model. Then we loom up the sentiment values for each word from

the sentiment lexicon, that has the all pre recorded. The classifier told the sentiment values of tweets. This process can be take place in three main steps.

i. Register for twitter API
ii. Install dependencies
iii. Write script for sentiment analysis

Twitter API is an application programming interface. It is the gateway that let user access some server's internal functionality. One can read or write tweets from own application using twitter API. In second step user need to install dependencies which required reading the text form authentic account and calculate the sentiment values. Then script writing is required. Currently python programming language is mostly used in script writing for machine learning concept. Through script writing sentiment analysis results can be calculated and presented in desired format. Figure 1 shows the entire activity of sentiment analysis through twitter data.

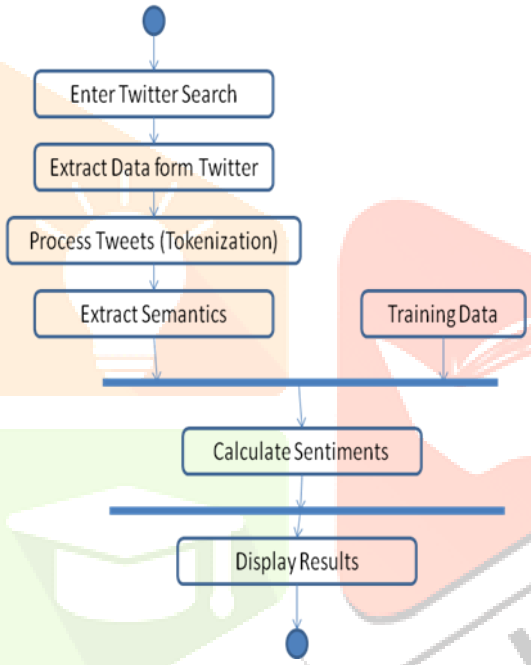

Figure 1: Activity diagram of sentiment analysis through twitter

5. Applications of Sentiment Analysis

[16]Sentiment analysis gives public opinion about any incident, product, person or topic. So that sentiment analysis has the huge area of applications. Some of them are included in this paper.

5.1 Movie Review

Many people give their reviews on movies through websites as well as on social media. A sentiment analysis system may be used to get a quick review of any movie. V.K. Singh et al. [17] presented experimental work on a domain specific feature-based heuristic for aspect-level sentiment analysis of movie reviews. Through this research they have calculated the sentiment level by analyzing textual content of movie reviews. Aggregated score of multiple review can be calculated to present cumulative review results through sentiment analysis.

5.2 Public Company Stock Analysis

[19]Stock value of any public company is mostly affected by public sentiments. Depending on the investor type company has, company may find value in the sentiment analysis in a few ways. Investors can do sentiment analysis before investing in the

company. Investors may find the current situation of company and also future prospects with company by sentiment analysis. It is use full for active traders to keep track of sentiment if they tend to trade in many companies simultaneously. Sentiment is often most indicative of price in the short term. It is very hard to keeping up update about many companies through news or any other medium. Sentiment analysis system can do it very affectively and fast for traders.

5.3 Political Topic Sentiments

[18]Politicians can use sentiment analysis to see how their actions and choices affect their image in the public. People can also use the political topic sentiment to see not only what is important to the public, but also how they currently feel about it. Political parts and public can take sentiments of other peoples about any politicians, that how much that politician is effective in particular area.

5.4 Geographic Sentiment Analysis

[20]This is a very interesting way to see what the general public of the world is talking about, and where about any issue. Often times, some time people do not get exact view of any incident happen in any other geographic are like in any other city, town, state or country. Sentiment Analysis system can help to know about this type of information.

5.5 Government Police Review

[21]Government of any state or country creates polices for the public. It is necessary to analyze public opinion about government policies. Sentiment analysis system can be used to analyze and review this policies and government can change according to public opinion.

5.6 Product Market Review

[22]Whenever any company launches their product in market, company required feed back or review of product to improve the quality of product. For that they use many review and feed bask systems on their website or through e-mail feedback system. But this process required some extra 3ffort and take time to get review. Instead of this time consuming process, company can use opinion mining system to get customer opinion in summarized form.

5.7 Recommendation System

[24]Recommendation systems are used to propose a product or services to customers based on their interest and requirement. To make any effective recommendation system sentiment analysis can be the part of that recommendation system. This helps to analyze the interest of customers about services or product.

5.8 Advertisement

World Wide Web is big platform for advertisement. Most of the public websites have reserved area for advertisements. To target appropriate audience f0r advertisement a mining system can be used. So that advertisements will be more effective.

5.9 Patients' opinion analysis in e-health system

[23]Barriers to use health connected quality of life activity systems embrace the time required to finish the forms and therefore the want for workers to be trained to grasp the results. a perfect system of health assessment must be clinically helpful, timely, sensitive to alter, culturally sensitive, low burden, low cost, involving for the patient and engineered into commonplace procedures. A replacement generation of short and easy-to-use tools to watch patient outcomes on an everyday basis can be implemented. By the use of sentiment analysis tool this can be achieve easily.

6. Challenges in Sentiment Analysis System

[25]It is nearly impossible to create a perfect sentiment analysis system for content available on social media platform. It has been observed only 80% accuracy can be archive through sentiment analysis system. In this section some of major challenges faced by researchers are included.

Quality of Content

[26]Social media like Tweeter, Facebook or Instagram provide a open platform to express user views. Thousands of people give their view on any topic, weather they belongs to that filed or not. In that condition no one can sure about the accuracy of content. Some time fake is also posed by people on social media. Users also commit typing mistakes freely on social media. So quality of content is always problem with social media.

6.1 Language Problem

[27]Present social media platforms are multi language sported. People share their views in many languages. Each communication language has its own set of vocabulary and grammar. So it is always a big challenge to develop a sentiment analysis system which can extracts sentiments from different types of language content simultaneously.

6.2 Geographical Area

[28]Sentiments about any incident or issue may be affected by geographical area. It might be possible that peoples of different geographical area think about differently on same issue. So geographical area might require some consideration while development of sentiment analysis system.

7. Conclusion

By this research it has been observed that sentiment analysis system is a very effective system to get quick review of any issue or incident. This analysis helps us to take important decision. As the application section of this paper shows the various areas where this sentiment analysis system can be used like in movie review, marketing, recommendation system, political issues and etc. It has been also found by the study that manly two types of sentiment analysis system are there. One is based on good and bad word frequency count, which is known as “Bag of Word” model and another next one is based on semantic of text, in this model machine learning concepts are used. Text based sentiment analysis system can be affected by quality of content, geographical area and language of text. It has been observed that there are lots of research scope in this filed to get a perfect sentiment analysis system

References

[1] Walaa Medhat , Ahmed Hassan , Hoda Korashy, "Sentiment analysis algorithms and applications:A survey", Ain Shams Engineering Journal, (2014) 5, pp. 1093–1113.

[2]. Ko Youngjoong, Seo Jungyun, “Automatic text categorization by unsupervised learning. In: Proceedings of COLING-00”, the 8th international conference on computational linguistics; 2000.

[3] Kang Hanhoon, Yoo Seong Joon, Han Dongil, “Senti-lexicon and improved Naı¨ve Bayes algorithms for sentiment analysis of restaurant reviews”. Expert Syst Appl 2012;39:6000–10.

[4] Ortigosa-Herna´ ndez Jonathan, Rodrı´guez Juan Diego, Alzate Leandro, Lucania Manuel, Inza In˜ aki, Lozano Jose A. “Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers.”, Neurocomputing 2012; 92:98–115.

[5] Kaufmann JM. JMaxAlign, “A Maximum Entropy Parallel Sentence Alignment Tool”. In: Proceedings of COLING’12: Demonstration Papers, Mumbai; 2012. p. 277–88.

[6] Chin Chen Chien, Tseng You-De. “Quality evaluation of product reviews using an information quality framework.” Decis Support Syst 2011;50:755–68.

[7] Ruiz M, Srinivasan P. "Hierarchical neural networks for text categorization." In: Presented at the ACM SIGIR conference; 1999.

[8] Hu Minging, Liu Bing. "Mining and summarizing customer reviews." In: Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD'04) 2004.

[9] Hatzivassiloglou V, McKeown K. "Predicting the semantic orientation of adjectives." In: Proceedings of annual meeting of the Association for Computational Linguistics (ACL'97); 1997.

[10] Moreo A, Romero M, Castro JL, Zurita JM. "Lexicon-based comments-oriented news sentiment analyzer system." Expert Syst Appl 2012;39:9166–80.

[11] Wille R. "Restructuring lattice theory: an approach based on hierarchies of concepts." In: I. Rival, Reidel, Dordrecht-Boston; 1982, p. 445–70.

[12] Li S, Tsai F. "Noise control in document classification based on fuzzy formal concept analysis." In: Presented at the IEEE International Conference on Fuzzy Systems (FUZZ); 2011.

[13] Mudinas Andrius, Zhang Dell, Levene Mark, "Combining lexicon and learning based approaches for concept-level sentiment analysis." Presented at the WISDOM'12, Beijing, China; 2012.

[14] Cambria Erik, Havasi Catherine, Hussain Amir. SenticNet 2: a semantic and affective resource for opinion mining and sentiment analysis. In: Proceedings of the twenty-fifth international florida artificial intelligence research society conference; 2012.

[15] Cambria Erik, Benson Tim, Eckl Chris, Hussain Amir. "Sentic PROMs: application of sentic computing to the development of a novel unified framework for measuring health-care quality.", Expert Syst Appl 2012;39:

[16] V. K. Singh , R. Piryani , A. Uddin , P. Waila, "Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification", International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 22-23 March 2013,

[17] Abinash Tripathya,* , Ankit Agrawalb , Santanu Kumar Rathc 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015) Classification of Sentimental Reviews Using Machine Learning Techniques

[18]MohammedElsaidMoussaEnsafHusseinMohamed Mohamed HassanHaggag "A survey on opinion summarization techniques for social media" Future Computing and informatics 2018

[19] Neurocomputing Volume 214, 19 November2016, *MohammadNoorInjadat[a]FadiSalo[a]Ali BouNassif[b]* "Data mining techniques in social media: A survey" 2016

[20]AdityaBhardwaj[a]YogendraNarayan[b]Vanraj[c]Pawan[a]MaitreyeeDuttaSentiment Analysis for Indian Stock Market Prediction Using Sensex and NiftyProcedia Computer Science Volume 70, 2015

[21]SoYeopYooJeInSong[1]OkRanJeongExpert Systems with Applications "Social Media Contents based Sentiment Analysis and Prediction System"28 March 2018

[22] Sefa ŞahinKoç[a]MertÖzer[b]İsmail HakkıToroslu[a]HasanDavulcu[b]JeremyJordan "Triadic co-clustering of users, issues and sentiments in political tweets" 2018

[23]Alvin OtiMensah[c]AriannaEstorelli[a] Elsevier *Expert Systems with Applications* A literature review for recommender systems techniques used in microblogs August 2018

[24] MillaSiikanen[a]KęstutisBaltakys[a]JuhoKanniainen[a]RaviVatrapu[bc]RaghavaMukkamala[bc]AbidHussain[b] Elsevier "Facebook drives behavior of passive households in stock markets"2018

[25] Mohammed ElsaidMoussaEnsaf HusseinMohamedMohamed HassanHaggag Elsevier"A survey on opinion summarization techniques for social media"2018

[26] MingxinGan[a]RuiJiang Elsevier "FLOWER: Fusing global and local associations towards personalized social recommendation"2018

[27] CharalamposKaryotis[a]FaiyazDoctor[d]RahatIqbal[a]AnneJames[b]VictorChang[c] Elsevier"A fuzzy computational model of emotion for cloud based sentiment analysis" 2018