IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

Sentiment Analysis and Rating Prediction Approaches - A survey

Dr. R.PRIYA[1], SNEHA. R [2]

Associate Professor & Head, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India [1]

M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India [2]

Abstract: The vast amount of web contents in the internet arise the demand for data mining techniques. To handle this content and huge data, in recent years, researchers have proposed various approaches with the aim to discover the essential information from the web. Product or service retrieval from the huge data is became very trouble, so users reviews and ratings are considered as a major point for recommendation and filtering. But, the user's reviews are very high and this need a lot of time to take decision from the reviews. So there is a need to summarize the reviews with finding sentiment analysis. This paper presents a survey on various Sentiment Analysis and mining techniques with different applications. From the comparative study, the new and optimal method can be detected and used for mobile app recommendation with Rating Prediction from huge set of text reviews.

Keywords: Data Mining, Text Mining, Opinion Analysis, Sentiment Analysis, Rating Prediction, Mobile App Recommendation

I. INTRODUCTION

In the current web scenario, web contents are growing drastically and huge in size. Analyzing and finding desired knowledge from the huge content is possible with the help of data mining. Data mining is an essential part of current applications like e-commerce, web search and others. In web search and e-commerce applications, the content recommendation is probably based on the ratings and popularity [1]. Sometimes the ratings are not explicitly given, thus it decreases the product or content reach. Opinion or sentiment analysis is trending in all type of social networks and e-commerce applications. Sentiment analysis is helpful for the product rating prediction based on the reviews.

This paper gathers a list of approaches and methods used to find the sentiment from the customers review. Sentiment analyses are vital process to most web recommendation activities because they are major influencers of the others interest. At the time decision making, user wishes to know other people opinions. In the current trend, businesses and organizations always want to find consumer or public opinions about their products and services [2]. This is not at all helpful to the business promoters; it also helps the user to know about the product or services which they are going to receive. Before buying or using any product/service, the users want to know the reviews about the product or service. But due to huge set of data, this is necessary to find the opinion about the product and predicts the rating of the product.

In common, sentiment analysis helps for gathering data regarding positive as well as negative features of the particular product/service. Finally, positive as well as excellent opinions acquired regarding a certain product are recommended to the customer. In order to promote marketing, large companies and business people make the use of opinion mining. To mine sentiments, the reviews collected can be analysed at three levels, such as document based, feature-based and sentence based.

Fig 1.0 shows the basic process of the sentiment analysis. This collects the reviews from different sources like face book, twitter, blogs and other social Medias.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

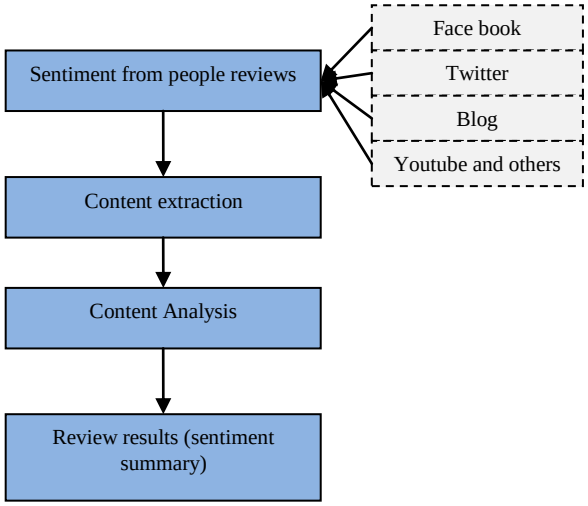

Fig 1.0 Process of Sentiment Analysis

Recently, online opinions have become a kind of virtual profit for business companies who are searching for ways to market their products, identify new trends as well as manage their positions [3]. However, the product ratings are given by the promoters to increase their business. Several organizations utilize sentiment mining systems for tracking consumer inputs in online shopping and review sites.

A. Applications of Sentiment Mining:

Most of the online shopping web site allows user to express their opinion on product. Customer can view review of product, they can compare their features. If review is mined and result is available in graphics format so it will be easy for customer to compare features [4]. The applications such as Internet Movie Database provides the review about the movie or television show, Government can mine the opinion on their public policies and rules, Customer review on product are useful for manufacturer and Research & Development. Department can improve the feature of product and students' opinion on e-learning system is helpful to improve service. These are the popular area of sentiment mining. For assisting customers, a particular trend is the creation of sites which automatically classifies and catalogue user reviews. An example of this type of web platform is recommendation sites, in which consumers can make an opinion about specific products they want to buy.

Organizations are also focused on how their image is reflected in customer opinion as well as market perception regarding already available products or during the release of new products. Hence, there is huge investment in research for the formulation of Business Intelligence (BI) systems which utilize semi-structured or unstructured data, apart from corporate data for market analyses [5]. Sentiment mining assists organizations in improving the consumer relationship management as well as in making studies regarding their attitudes toward the particular brand. The information obtained from web data assists them in effectuating estimations as well as helps managers in making informed decisions.

Another major application of sentiment mining is helping sociologists study people's sentiments. Conducting sociological studies involves the usage of considerable amount of funds in taking polls. To be precise, the polls are to be conducted in huge populations. Sentiment mining assists in obtaining results through considerably lesser amount of effort and is capable of targeting huge masses of the population. Determination of sentiments with regard to policies or the studying of effects of particular laws or even reforms on populations assists the governments in making decisions or in promoting policies agreed by the masses.

Furthermore, sentiment mining utilized in security systems assists in the increasing of accuracy in detection of spams in anti-spam systems. Argument mapping software assists in the organization logically those policy statements through exploitation of logical links between them. Voting Advise Applications

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

assists voters understanding which political party (or other voters) have nearer positions to theirs. Automatic content analysis assists in the processing of huge quantities of qualitative data. There are several tools available that combine statistical algorithm with semantics and ontologies, as well as machine learning with human supervision [6]. There are several other applications utilizes the sentiment. The applications are such as online message sentiment filtering, mail sentiments classification, web blog authors' attitudes analyses and so on.

B. Challenges in Sentiment Analysis:

Sentiment Analysis presents various difficulties. The first is sentiment words that are regarded as positive in one case and are regarded as negative in another. The second is that individuals do not always convey their sentiments in a similar manner. Almost all conventional text processing depends on the fact that minor variations between two sets of text do not alter the meaning to a great extent [7]. Named entities are definite noun phrases which refer to particular kinds of individuals, like companies, persons, dates and others. The aim of named entity extraction is the identification of all textual features in a text.

Information is present in various shapes as well as sizes. The complexity of natural language makes it very hard to access the information present in opinion pieces. The sentiment determination is a task that assigns a sentiment polarity to a word, sentence or document. A conventional way for sentiment polarity assignment is the usage of the sentiment lexicon. The adjectives of a sentence are particular significance in sentiment mining as they have the greater probability of carrying information while sentiment analysis issue is taken into consideration. Co-reference resolution is to be done in aspect level and entity level. These references must be effectively resolved for producing correct results. Relation extraction is the task of finding the syntactic relation between words in a sentence [8]. The semantics of a sentence can be found out by extracting relations between words and this can be done by knowing the word dependencies. This is also a major research area in NLP and serious researches are going on to solve this problem. A sentiment classifier that is trained to classify opinion polarities in a domain may produce miserable results when the same classifier is used in another domain. The sentiment is expressed differently in different domains. Sentiment analysis is a problem which has high domain dependency. Therefore cross-domain sentiment analysis is a challenging problem that has to be unfolded.

II. LITERATURE REVIEW

Machine learning techniques have been widely deployed for sentiment and opinion classification at various levels, e.g., from the document level, to the sentence and word/phrase level. On the document level, one tries to classify products as positive, negative, or neutral, based on the overall sentiments expressed by opinion holders. There are several lines of representative work at the early stage [9]. Lin [10] used weakly supervised learning with mutual information to predict the overall document sentiment and opinion by averaging out the sentiment and opinion orientation of phrases within a document. Pang et al. [11] classified the polarity of movie reviews with the traditional supervised machine learning approaches and achieved the best results using SVMs. In their subsequent work [12], the sentiment and opinion classification accuracy was further improved by employing a subjectivity detector and performing classification only on the subjective portions of reviews.

The annotated movie review data set (also known as polarity data set) used in [13] and [14] has later become a benchmark for many studies [15]. Whitelaw et al. [15] used SVMs to train on combinations of different types of appraisal group features and bag-of-words features, whereas Kennedy and Inkpen [16] leveraged two main sources, i.e., General Inquirer and choose the right word, and trained two different classifiers for the sentiment and opinion classification task. As opposed to the work [17] that only focused on sentiment and opinion classification in one particular domain, some researchers have addressed the problem of sentiment and opinion classification across domains. Aue and Gamon [18] explored various strategies for customizing sentiment and opinion classifiers to new domains, where training is based on a

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

small number of labelled examples and large amounts of unlabeled in-domain data. It was found that directly applying a classifier trained on a particular domain barely outperforms the baseline for another domain.

In the same way, more recent work focused on domain adaptation for sentiment and opinion classifiers. Blitzer et al. [14] addressed the domain transfer problem for sentiment and opinion classification using the Structural Correspondence Learning (SCL) algorithm, where the frequent words in both source and target domains were first selected as candidate pivot features and pivots were then chosen based on the mutual information between these candidate features and the source labels. They achieved an overall improvement of 46 percent over a baseline model without adaptation. Li and Zong [19] combined multiple single classifiers trained on individual domains using SVMs.

However, their approach relies on labelled data from all domains to train an integrated classifier and thus may lack flexibility to adapt the trained classifier to other domains where no label information is available.

All the aforementioned work shares some similar limitations such as, the earlier studies focused on sentiment and opinion classification alone without considering the mixture of topics in the text, which limits the effectiveness of the mining results to users. And most of the approaches favour supervised learning, requiring labelled corpora for training, and potentially limiting the applicability to other domains of interest. Compared to the traditional topic-based text classification, sentiment and opinion classification is deemed to be more challenging as sentiment and opinion is often embodied in subtle linguistic mechanisms such as the use of sarcasm or incorporated with highly domain-specific information. Among various efforts for improving sentiment and opinion detection accuracy, one of the directions is to incorporate prior information from the general sentiment and opinion lexicon (i.e., words bearing positive or negative sentiment) into sentiment and opinion models.

These general lists of sentiment and opinion lexicons can be acquired from domain independent sources in many different ways, i.e., from manually built appraisal groups [15], to semi automatically [16] or fully automatically [17] constructed lexicons. When incorporating lexical knowledge as prior information into a sentiment-topic model, and deevskaia and Bergler [20] integrated the lexicon-based and corpus-based approaches for sentence-level sentiment and opinion annotation across different domains. A recently proposed nonnegative matrix trifactorization approach [21] also employed lexical prior knowledge for semi-supervised sentiment and opinion classification, where the domain-independent prior knowledge was incorporated in conjunction with domain-dependent unlabeled data and a few labelled products. However, this approach performed worse than the existing model on the movie review data even with 40 percent labelled products.

ATOM performs sentiment and opinion detection in a mixture of topics simultaneously. Although work in this line is still relatively sparse, some studies have preserved a similar vision [22]. Most closely related to the work is the Topic-Sentiment and opinion Model (TSM) [23], which models mixture of topics and sentiment and opinion predictions for the entire document. However, there are several intrinsic differences between ATOM and TSM.

First, TSM is essentially based on the probabilistic latent semantic indexing (pLSI) [24] model with an extra background component and two additional sentiment and opinion subtopics, whereas ATOM is based on LDA, Second, regarding topic extraction, TSM samples a word from the background component model if the word is a common English word. Otherwise, a word is sampled from either a topical model or one of the sentiment and opinion models (i.e., positive or negative sentiment and opinion model). Thus, in TSM the word generation for positive or negative sentiment and opinion is not conditioned on topic. This is a crucial difference compared to the ATOM model as in ATOM one draws a word from the distribution over words jointly conditioned on both topic and sentiment and opinion label.

Third, for sentiment and opinion detection, TSM requires post processing to calculate the sentiment and opinion coverage of a document, while in ATOM the document sentiment and opinion can be directly

obtained from the probability distribution of sentiment and opinion label given a document. Other models by Titov and McDonald are also closely related to the work, since they are all based on LDA. The Multi-Grain Latent Dirichlet Allocation model (MG-LDA) is argued to be more appropriate to build topics that are representative of predictable aspects of customer reviews, by allowing terms being generated from either a global topic or a local topic. Being aware of the limitation that MG-LDA is still purely topic-based without considering the associations between topics and sentiments, Titov and McDonald further proposed the Multi-Aspect Sentiment and opinion model (MAS) by extending the MG-LDA framework. The major improvement of MAS is that it can aggregate sentiment and opinion text for the sentiment and opinion summary of each rating aspect extracted from MG-LDA. The proposed model differs from MAS in several aspects.

First, MAS works in a supervised setting as it requires that every aspect is rated at least in some products, which is infeasible in real-world applications. In contrast, ATOM is weakly supervised with only minimum prior information being incorporated, which in turn is more flexible. Second, the MAS model was designed for sentiment and opinion text extraction or aggregation, whereas ATOM is more suitable for the sentiment and opinion classification task.

III. PROBLEM DEFINITION

There are several challenges in analyzing the sentiment of the web user reviews. First, a word that is considered to be positive in one situation may be considered negative in another situation. Take the word "long" for instance. If a customer said a laptop's battery life was long, that would be a positive opinion.

If the customer said that the laptop's start-up time was long, however, that would be a negative opinion [25]. These differences mean that an opinion system trained to gather opinions on one type of product or product feature may not perform very well on another. Another challenge would arise because the people don't always express opinions the same way. Most traditional text processing relies on the fact that small differences between two pieces of text don't change the meaning very much. In opinion mining, however, "the movie was great" is very different from "the movie was not great". People can be contradictory in their statements. Most reviews will have both positive and negative comments, which is somewhat manageable by analyzing sentences one at a time.

However, the more informal the medium (twitter or blogs for example), the more likely people are to combine different opinions in the same sentence. For example: "the movie bombed even though the lead actor rocked it" is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context. For example, "That movie was as good as his last one" is entirely dependent on what the person expressing the opinion thought of the previous film.

The recent IEDR work does not currently extract non-noun opinion features due to the limitation of only considering nouns (noun phrases) for candidate feature extraction in the dependence parsing phase.

CONCLUSION

The survey demonstrated techniques and approaches already in research to organize and summarize various feedback or reviews using text mining techniques. The most of the existing approaches to sentiment classification is based on supervised learning, showed in the existing system with and show models target sentiment, opinion and rating detection simultaneously in a semi supervised fashion. For general domain sentiment classification, by incorporating a small amount of domain independent prior knowledge, the survey shows, the document level sentiment detection model achieved the better or comparable performance compared to existing semi-supervised approaches despite using un-labeled data's, which demonstrates the flexibility of dataset in the task.

REFERENCES

[1]. Li, Yu, Liu Lu, and Li Xuefeng. "A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce." *Expert systems with applications* 28.1 (2005): 67-77.
[2]. Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5.1 (2012): 1-167.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

[3]. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and Trends® in Information Retrieval* 2.1–2 (2008): 1-135.
[4]. Meshram, Milind D. "Feature based opinion mining: an overview." *Proceedings of National Conference on Emerging Trends: Innovations and Challenges in IT*. Vol. 19. 2013.
[5]. Bucur, Cristian. "Opinion Mining platform for Intelligence in business." *Economic Insights--Trends and Challenges* 66.3 (2014).
[6]. Osimo, David, and Francesco Mureddu. "Research challenge on opinion mining and sentiment analysis." *Universite de Paris-Sud, Laboratoire LIMSI-CNRS, Bâtiment* 508 (2012).
[7]. Vinodhini, G., and R. M. Chandrasekaran. "Sentiment analysis and opinion mining: a survey." *International Journal* 2.6 (2012): 282-292.
[8]. Varghese, Raisa, and M. Jayasree. "A survey on sentiment analysis and opinion mining." *International Journal of Research in Engineering and Technology* 2.11 (2013): 312-317.
[9]. Kaur, Amandeep, and Vishal Gupta. "A survey on sentiment analysis and opinion mining techniques." *Journal of Emerging Technologies in Web Intelligence* 5.4 (2013): 367-371.
[10]. Lin, Chenghua, et al. "Weakly supervised joint sentiment-topic detection from text." *IEEE Transactions on Knowledge and Data engineering* 24.6 (2012): 1134-1145.
[11]. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." *Foundations and Trends® in Information Retrieval* 2.1–2 (2008): 1-135.
[12]. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.
[13]. Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003.
[14]. Blitzer, John, Mark Dredze, and Fernando Pereira. "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification." *ACL*. Vol. 7. 2007.
[15]. Whitelaw, Casey, Navendu Garg, and Shlomo Argamon. "Using appraisal groups for sentiment analysis." *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005.
[16]. Kennedy, Alistair, and Diana Inkpen. "Sentiment classification of movie reviews using contextual valence shifters." *Computational intelligence* 22.2 (2006): 110-125.
[17]. Ye, Qiang, Ziqiong Zhang, and Rob Law. "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches." *Expert systems with applications* 36.3 (2009): 6527-6535.
[18]. Aue, Anthony, and Michael Gamon. "Customizing sentiment classifiers to new domains: A case study." *Proceedings of recent advances in natural language processing (RANLP)*. Vol. 1. No. 1-3. 2005.
[19]. Li, Shoushan, and Chengqing Zong. "Multi-domain sentiment classification." *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*. Association for Computational Linguistics, 2008.
[20]. Andreevskaia, Alina, and Sabine Bergler. "Mining WordNet for a Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses." *EACL*. Vol. 6. 2006.
[21]. Li, Tao, Yi Zhang, and Vikas Sindhwani. "A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge." *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, 2009.
[22]. Kanayama, Hiroshi, and Tetsuya Nasukawa. "Fully automatic lexicon expansion for domain-oriented sentiment analysis." *Proceedings of the 2006 conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2006.
[23]. Mei, Qiaozhu, et al. "Topic sentiment mixture: modeling facets and opinions in weblogs." *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007.
[24]. Hofmann, Thomas. "Probabilistic latent semantic indexing." *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999.
[25]. Ding, Xiaowen, and Bing Liu. "The utility of linguistic rules in opinion mining." *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007.