ISSN: 2321-7782 (Online)
Volume 2, Issue 2, February 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Paper / Case Study
Available online at: www.ijarcsms.com

Review Paper on Sentiment Analysis is – Big Challenge

Siddhi Patni[1]
Department of Computer Science and Engineering
GH Raisoni College of Engineering And Management
Amravati – India

Avinash Wadhe[2]
Prof.
GH Raisoni College of Engineering And Management
Amravati – India

Abstract: ***Our day-to-day life has always been influenced by what people think. Ideas and opinions of others have always affected our own opinions. As the Web plays an increasingly significant role in people's social lives, it contains more and more information concerning their opinions and sentiments. The distillation of knowledge from this huge amount of unstructured information, also known as opinion mining and sentiment analysis. Nowadays, with the rapid evolution of smart phones, mobile applications (Mobile Apps) have become essential parts of our lives. However, it is difficult for consumers to keep track and understand the app sphere because new apps are entering market every day. Such a large amount of apps seems to be a great opportunity for customers to buy from a wide selection range. But, first they have to understand what the apps do, how are they viewed by other consumers and then they have to purchase the apps to use on their smart phones.. It is very challenging for a potential user to read all of them to make a decision. Also, app developers have difficulties in finding out how to improve the app performance based on overall ratings alone and would benefit from understanding the thousands of textual comments. Because the identification of sentiment is often exploited for detecting polarity, however, the two fields are usually combined under the same umbrella or even used as synonyms. Both fields use data mining and natural language processing (NLP) techniques to discover, retrieve, and distil information and opinions from the World Wide Web's vast textual information.***

Keywords: Sentiment Analysis, NLP approach, Machine Learning, Android Apps.

I. INTRODUCTION

With the rapid evolution of smart phones, mobile applications (Mobile Apps) have become essential parts of our lives. Zynga Game Network and Rovio Entertainment are examples of companies that gained huge game apps market shares .However, it is difficult for consumers to keep track and understand the app sphere because new apps are entering market every day. It is reported that Android market reached half a million apps in September 2011[10]. As of October, 2012, 0.675 million Android apps are available on Google Play App Store [11]. Such a large amount of apps seems to be a great opportunity for customers to buy from a wide selection range. But, first they have to understand what the apps do, how are they viewed by other consumers and then they have to purchase the apps to use on their smart phones. According to a recent behavior survey, 98 per cent of shoppers consider online customer reviews as a major purchase decision factor [9]. We believe mobile app users also consider online app reviews as a major influence for paid apps. Typically, online customer reviews contain two parts, ratings and textual comments. Rating indicates the overall evaluation of customer experiences using a numeric scale, but textual comments are capable of telling more insightful stories that the overall ratings cannot. After few months of a new app launched in the market, there could be over ten thousand textual comments from users. It is very challenging for a potential user to read all of them to make a decision. Also, app developers have difficulties in finding out how to improve the app performance based on overall ratings alone and would benefit from understanding the thousands of textual comments.

II. LITERATURE SURVEY

Erik Cambria, BjörnSchuller, Catherine Havasi [3] this paper describes new avenues in sentiment analysis. The Web has changed from "read- only" to "read-write." This evolution created enthusiastic users interacting and sharing through social networks, online communities, blogs, wikis, and other collaborative media. Collective knowledge has spread throughout the web, particularly in areas related to everyday life, such as e-commerce, tourism, education, and health. Engineers and computer scientists use machine-learning techniques for automatic affect classification from video, voice, text, and physiology. Psychologists combine the long tradition of emotion research with their own discourse, models, and methods. Future opinion-mining systems need broader and deeper common and commonsense knowledge bases. LisetteGarcía-Moya, Henry Anaya-Sánchez, and Rafael Berlanga-Llavori, [1] In this paper author focuses on a new methodology based on language models retrieves product features and opinions from a collection of free-text customer reviews about products and services. Changbo Wang, Zhao Xiao, Yuhua Liu, YanruXu, Aoying Zhou, and Kang Zhang [2] this paper has introduced a new visualization system for analyzing, visualizing and verifying the sentiments of Web users on public topics. A text-based sentiment mining method and a model-driven prediction approach have been used to analyze the public sentiments on hot topics.

Siddhi Patni, Avinash Wadhe [7] in this paper major tasks, various challenges, and applications of sentiment analysis. Most work has been done on product reviews – documents that have a definite topic. More general writing with varied domains, such as blog posts, tweets, posts and web pages, have recently been creating & receiving attention. Future work in expanding existing techniques to handle more general writings and crossing domains is an exciting opportunity for both academia and businesses .Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou [6] in this paper sentiment classification is applied to the movie reviews, and rating information is based on sentiment-classification results. In feature-based summarization, product-feature identification plays an essential role, a novel approach is proposed based on LSA to identify related product features. Product features and opinion words will be used as the basis for feature-based summarization. The number of features plays an important role in SVM-model to classify the reviews.

Raymond Hsu, Bozhi See, Alan Wu [12] in this paper the use of colloquial and slang language in most of the confessions. The use of spell checking corrected for this somewhat. Nonetheless, the synset and sentiment lexicons we used are better suited to more formal styles of writing. An alternative approach is to replace our synsets and lexicons with "slang" versions or even the automatic generation of sentiment lexicons on a slang corpus. Another area of interest is the difficulty in correlating topics with sentiment. Intuition says that topics themselves should portray different sentiments, and so should be useful for sentiment analysis. This method turns out to be fairly crude, as sometimes topics may be too neutral or too general to actually be good indicators of mood. For example, one of the topics found with LDA turned out to contain the topic about relation- ships. It is possible for someone to complain angrily about their current relationship, cry over the impending end of a relationship, or laugh because of a happy moment during the relationship. All of these get mapped into the same topic, but each has a substantially different mood.

G.Vinodhini, RM. Chandrasekaran [8] In this paper it is found that sentiment detection has a wide variety of applications in information systems, including classifying reviews, summarizing review and other real time applications. It is found that sentiment classifiers are severely dependent on domains or topics and also found that different types of features and classification algorithms are combined in an efficient way in order to overcome their individual drawbacks and benefit from each other's merits, and finally enhance the sentiment classification performance. In future, more work is needed on further improving the performance measures. Sentiment analysis can be applied for new applications. Although the techniques and algorithms used for sentiment analysis are advancing fast, however, a lot of problems in this field of study remain unsolved. The main challenging aspects exist in use of other languages, dealing with negation expressions; produce a summary of opinions based on product features/attributes, complexity of sentence/ document, handling of implicit product features, etc. More future research could be dedicated to these challenges.

Jiawen Liu, Mantosh Kumar Sarkar and Goutam Chakraborty [5] in this paper for android apps, NLP rule-based models is carefully designed in SAS® Sentiment Analysis Studio 12.1 for predicting sentiments in test data. The current default versions for statistical models in SAS® Sentiment studio do not allow for much customization. This may have contributed to the poorer performance of statistical models than NLP models in this study. The NLP rule based models also provide deeper insights than statistical models in understanding consumers' sentiments. For example, we find that app users are very addicted to the game app, but not happy for being charged more as they play more; they are pleased with graphics design of the widget app, but not ok with the app accessing their personal information. No spelling check and no stop list were used in this study. These could be considered in future research to perhaps get better text mining and sentiment mining results.

Albert Weichselbraun, Stefan Gindl and Arno Scharl[4] In this paper a method is used to improve sentiment analysis by using contextualized sentiment lexicons is to disambiguate sentiment terms. A graph-based component for concept identification refines these lexicons and uses WordNet to ground ambiguous sentiment terms to concepts. This grounding process provides a clear distinction between positive and negative concepts, and paves the way for incorporating semantic databases into the sentiment analysis process. Changbo Wang, Zhao Xiao, Yuhua Liu, YanruXu, Aoying Zhou, and Kang Zhang [2] in this paper a text-based sentiment mining method and a model-driven prediction approach is used to analyse the public sentiments on hot topic. SentiView can be used to analyse and visualize mass Web information effectively in many applications. SentiView builds upon and extend several ideas from state-of- the-art techniques to enable advanced visual analysis of public sentiments on popular topics on the Internet. Three system components designed in SentiView are showing rich information from different aspects at once and provide flexibility for varying task. Lisette García-Moya, Henry Anaya-Sánchez, and Rafael Berlanga-Llavor[1] in this paper a new methodology for the retrieval of product features from a collection of customer reviews about a product or service is that it doesn't require any training set of product features, and over several collections of customer reviews in English. For future work, we plan to integrate our models into a probabilistic topic-modelling framework and a plan to extend our methodology to model the polarity of the opinion words ascribed to product features.

III. DATA SOURCE FOR REVIEW COLLECTION

User's opinion is a major criterion for the improvement of the quality of services rendered and enhancement of the deliverables. Blogs, review sites, data and micro blogs provide a good understanding of the reception level of the products and services.

A. REVIEW SITES

A review site is a website which allows users to post reviews which give a critical opinion about people, businesses, products, or services. Most sentiment analysis work has been done on movie and product review sites [14, 15]. The purpose of a review is to appraise a specific object, thus it is a single domain problem.

B. BLOGS

The term web-log or blog refers to a simple webpage consisting of brief paragraphs of opinion, information, personal diary entries, or links, called posts, arranged chronologically with the most recent first, in the style of an online journal [13]. Sentiment analysis on blogs [16] has been used to predict movie sales, political mood and sales analysis.

C. FORUMS

Forums or message boards allow its members to hold conversations by posting on the site. Forums are generally dedicated to a topic and thus using forums as a database allows us to do sentiment analysis in a single domain.

D. SOCIAL NETWORKS

Social networking is online services or sites which try to emulate social relationships amongst people who know each other or share a common interest. Social networking sites allow users to share ideas, activities, events, and interests within their individual networks.

E. GOOGLE PLAY ANDROID APP STORE

Google Play Android App Store has a large and varied collection of Android Apps with rankings and user reviews. It extracted textual reviews having rich content from the App Store site [5].

IV. REVIEW ANALYSIS TECHNIQUES

Various techniques can be used in analysis of reviews.

A. PARSER

In order to refine data and improve the feature set, remove all HTML tags using a parser. This is essential towards refining the dataset because HTML tags do not convey emotions and would skew the feature vector by including phrases that have no se- mantic meaning (e.g. ' '). Emoticons, on the other hand, are an excellent way of conveying emotions through text because it captures the emotion of the writer by including a facial expression. Therefore, capturing this unique feature set and used it to improve the feature vector.

B. SPELL CHECKING

One of the issues is overfitting. There were many spelling errors in the text or the reviews that people give. In order to reduce problems of overfitting as a result of having too many unique spellings, a spell checker is used to correct all the spelling errors.

C. FEATURES

There are three features in our model: bag of words, WordNet2 synsets, and sentiment lexicons.

1. *Bag Of Words (Bow):* The BoW model is the most basic feature model in sentiment analysis. It treats each unique word token as a separate feature.

2. Wordnet Synsets: In order to further improve the quality of the feature set and decrease overfitting , WordNet is used to map the words in the confessions onto their synonym set (synset). By mapping words into their synset, the assumption can be made that the words of similar meaning elicit similar emotions. This reduces the number of unique features and also improves the coverage of each feature. This technique also allows handling words that do not occur in the training data if they hap- pen to be in the same synset as words that do occur in the training data.

3. *Sentiment Lexicons:* Sentiment lexicons are groupings of words into emotion and content categories, they improved the performance. It can be used for replacing the original words with their sentiment lexicon category. One of the sentiment lexicon used was Language Inquiry and Word Count (LIWC), a hand- engineered set of words and categories used by psychologists to group words by similar emotional and subject content. Also a feature, which also categorizes words by emotional and subject content, can be used. Like LIWC, the Harvard Inquirer was also hand-engineered by psychologists for the purpose of analyzing text. Both lexicons have been used in previous work on sentiment analysis.

4. *Stop Words:* Not surprisingly, function words such as 'and', 'the', 'he', 'she' occur very often across all confessions. Therefore, it makes little sense to put a lot of weight on such words when using bag of words to classify the documents. One common approach is to remove all words found in a list of high frequency stop words. A better approach is to consider each word's Term Frequency-Inverse Document Frequency (TF-IDF) weight. The intuition is that a frequent word that appears in

only a few confessions conveys a lot of information, while an infrequent word that appears in many confessions conveys very little in formation [12].

V. SENTIMENT ANALYSIS OPERATIONS

A. NATURAL LANGUAGE PROCESSING

Part-of speech (POS) tagging is often the most time consuming and challenging task before doing sentiment analysis of any text documents. Online textual reviews are often short, non-grammar sentences and contain slangs, abbreviations, and symbols which make the POS tagging even more difficult. For example, consider the following statement. "The game is good. I love its graphics design and I can play it for hours." In this comment, "game" is tagged as product and "graphics design" is tagged as feature. Products and features are tagged as nouns. We can define the synonym list of products and features. This feature can be because of uncertain and non-grammar online reviews. For example, consider the following comment. "I love the high res". Here "res" likely refers to resolution, and resolution is a word which is similar to graphics. One of the difficulties that researchers often face in doing sentiment analysis is with textual reviews that contain mixture sentiments such as the following comment. "I love the graphics, but it drains battery a lot". Because we are doing feature based sentiment analysis, we are able to easily handle such reviews. In this case, the sentiment is positive on "graphics/design" and negative on "battery" [5]. For this CLASSIFIER, CONCEPT, CONCEPT_RULE, and PREDICATE_RULE rules can be used. CLASSIFIER rules are used to match the words which can be used only for a feature. For example, "expensive" can only be used for feature "price". CONCEPT rules are used to locate related terms [17]. We primarily used "@" for this rule. Symbol "@"means that it matches all noun and verb forms of a word [17].

B. MACHINE LEARNING APPROACH

The machine learning approach applicable to sentiment analysis mostly belongs to supervised classification in general and text classification techniques in particular. In a machine learning based classification, two sets of documents are required: training and a test set. A training set is used by an automatic classifier to learn the differentiating characteristics of documents, and a test set is used to validate the performance of the automatic classifier. A number of machine learning techniques have been adopted to classify the reviews. Machine learning techniques like Naive Bayes (NB), maximum entropy (ME), and support vector machines (SVM) have achieved great success in text categorization. The other most well- known machine learning methods in the natural language processing area are K-Nearest neighborhood, ID3, C5, centroid classifier, winnow classifier, and the N-gram model.

1. Naive Bayes: It is a simple but effective classification algorithm. The Naive Bayes algorithm is widely used algorithm for document classification .The basic idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The naive part of such a model is the assumption of word independence. The simplicity of this assumption makes the computation of Naive Bayes classifier far more efficient [8].

2. Support vector machines (SVM): It is a discriminative classifier is considered the best text classification method. The support vector machine is a statistical classification method proposed by Vapni. Based on the structural risk minimization principle from the computational learning theory, SVM seeks a decision surface to separate the training data points into two classes and makes decisions based on the support vectors that are selected as the only effective elements in the training set. Multiple variants of SVM have been developed in which Multi class SVM is used for Sentiment classification. The idea behind the centroid classification algorithm is extremely simple and straightforward. Initially the prototype vector or centroid vector for each training class is calculated, then the similarity between a testing document to all centroid is computed, finally based on these similarities, document is assigned to the class corresponding to the most similar centroid.

3. K-nearest neighbor (KNN): It is a typical example based classifier that does not build an explicit, declarative representation of the category, but relies on the category labels attached to the training documents similar to the test document. Given a test

document d, the system finds the k nearest neighbors among training documents. The similarity score of each nearest neighbor document to the test document is used as the weight of the classes of the neighbor document.

4. Winnow: It is a well-known online mistaken-driven method. It works by updating its weights in a sequence of trials. On each trial, it first makes a prediction for one document and then receives feedback; if a mistake is made, it updates its weight vector using the document. During the training phase, with a collection of training data, this process is repeated several times by iterating on the data.

VI. APPLICATION

The applications are broadly classified into the following categories.

A. APPLICATIONS TO REVIEW-RELATED WEBSITES

Movie Reviews, Product Reviews *etc.*

B. APPLICATIONS AS A SUB-COMPONENT TECHNOLOGY

Detecting antagonistic, heated language in mails, spam detection, context sensitive information detection etc.

C. *APPLICATION IN BUSINESS AND GOVERNMENT INTELLIGENCE*

Knowing Consumer attitudes and trends

D. APPLICATIONS ACROSS DIFFERENT DOMAINS

Knowing public opinions for political leaders or their notions about rules and regulations in place etc.

VII. CONCLUSION

This paper illustrates the research area of Sentiment Analysis on reviews on product like android apps and its latest advances. It affirms the major tasks, various challenges, and applications of sentiment analysis. Most work has been done on product reviews – documents that have a definite topic. More general writing with varied domains, such as blog posts, tweets, posts and web pages, have recently been creating & receiving attention. Study of various machine learning technique and NLP technique are used. There is major advantage of sentiment analysis in mobile environment of analyzing the reviews of users using apps. For future work, we would like to explore its role in clustering and classification of reviews based on their sentiments.

References

1. LisetteGarcía-Moya, Henry Anaya-Sánchez, and Rafael Berlanga-Llavori, “Retrieving Product Features and Opinions from Customer Reviews”, IEEE 2013.
2. Changbo Wang, Zhao Xiao, Yuhua Liu, YanruXu, Aoying Zhou, and Kang Zhang,”SentiView: Sentiment Analysis and Visualization for Internet Popular Topics”, IEEE November 2013.
3. Erik Cambria, BjörnSchuller, Catherine Havasi, “New Avenues in Opinion Mining and Sentiment Analysis”, IEEE 2013.
4. Albert Weichselbraun, Stefan Gindl and Arno Scharl “Extracting and Grounding Contextualized Sentiment Lexicons”, IEEE 2013.
5. Jiawen Liu, Mantosh Kumar Sarkar and GoutamChakraborty, “Feature-based Sentiment Analysis on Android App Reviews Using SAS® Text Miner and SAS® Sentiment Analysis Studio”, SAS Global Forum 2013.
6. Chien-Liang Liu, Wen-Hoar Hsaio, Chia-Hoang Lee, Gen-Chi Lu, and Emery Jou” Movie Rating and Review Summarization in Mobile Environment”, IEEE MAY 2012.
7. Siddhi Patni ,AvinashWadhe , “Review paper on sentiment analysis using web 2.0 by classification method “,IJARCS 2012.
8. G. Vinodhini, RM. Chandrasekaran, “Sentiment Analysis and Opinion Mining”: A Survey ,IJARCSSE Volume 2, Issue 6, June 2012.
9. Ping! Zine Editor. “Pinterest Users Buy More Items Than Facebook Users, According to Survey”. Ping! Zine Web Tech Magazine. May, 2012.
10. Mikalajunaite, Egle. “Android Market reaches half a million successful submissions". Research2Guidiance. Oct, 2011.
11. Kincaid, Jason. “Android Market: 10 Billion Apps Served So Far, And Another 1 Billion Each Month”. Techcrunch. Dec, 2011.
12. Raymond Hsu, Bozhi See, Alan Wu,’’Machine Learning for Sentiment Analysis on the Experience Project”, 2010.

13. Anderson, P. What is Web 2.0? Ideas, technologies and implications for education. Technical report, JISC, 2007.
14. Theresa Wilson, JanyceWiebe, and Rebecca Hwa. Just how mad are you? Finding strong and weak opinion clauses. In Proceedings of AAAI, 2004, pages 761–769.
15. Dave K., Lawrence S, and Pennock D.M. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (WWW), 2003, pp.:519–528.
16. Mishne G. and Glance N. Predicting movie sales from blogger sentiment. In AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW), 2006: 155–158.
17. SAS Institute Inc. 2011. SAS® Sentiment Analysis Studio 1.3: User's Guide. Cary, NC: SAS Institute Inc.

AUTHOR(S) PROFILE

Miss. Siddhi S. Patni is doing M.E (CSE) from G.H Raisoni College of Engineering and Management, Amravati at Sant Gadge Baba Amravati University, Amravati, Maharashtra, India, and has done B.E in Information Technology from SGBAU, Amravati.

Prof. Avinash P. Wadhe: Received the B.E and from SGBAU Amravati university and M-Tech (CSE) From G.H Raisoni College of Engineering, Nagpur (an Autonomous Institute). He is currently an assistant Professor with the G.H Raisoni College of Engineering and Management, Amravati SGBAU Amravati University. His research interest include Network Security, Data mining and Fuzzy system .He has contributed to more than 20 research paper. He had awarded with young investigator award in international conference.

