The International Journal Of Engineering And Science (IJES)
||Volume||2 ||Issue|| 12||Pages|| 11-15||2013||
ISSN(e): 2319 – 1813 ISSN(p): 2319 – 1805

A Survey of Classification Methods and Applications for Sentiment Analysis

[1]M.Govindarajan , [2,]Romina M.
[1,] *Assistant Professor*
[2,] *JRF, DST-SERB, Fast Track Scheme for Young Scientist,*
[1,2,] *Department of Computer Science & Engineering Annamalai University Annamalai Nagar – 608002 Tamil Nadu, India*

----------ABSTRACT----------
The amount of data readily available is far beyond our capacity to analyze and understand. The internet revolution has added to this problem by having billions of customer's review data in its repositories. This has provoked an interest in sentiment analysis and opinion mining in the recent years. People have to rely on machines to classify and process the data as there are terabytes of review data available for a single product. In order to predict the customer sentiments it is very important to analyze the reviews as it not only helps in increasing profits but also goes a long way in improving and bringing out better products. The main problem in this sentiment analysis understands the use of negation and the classification of positive and negative sentiments recorded by the users as the syntax and semantics of the language varies according to the dialect, it is difficult to categorize an opinion. This paper presents a survey regarding the presently available techniques, applications and problems that appear in the field of opinion mining.

KEY WORDS: *sentiment mining, negation, syntax and semantics*

Date of Submission: 29 November 2013 ⟺ Date of Acceptance: 20 December 2013

I. INTRODUCTION

Sentiment or opinion mining refers to the type of natural language processing used to understand the moods, opinions and sentiments of the public regarding a particular product or a movie or an event. The availability of large amounts of data and the human tendency to always manipulate what other people think has been influential in a decision making process. This unique feature plays a vital role in deciding on matters that have financial, medical, social or other implications. Seeking second or third or many more opinions have fuelled the interest of researchers in the field of sentiment mining. With multiple reviews available for a single product and the enormous growth in the number of internet users it has become indispensable to develop a system that collects, builds, analyzes, and classifies the comments or a review posted online.

Usually these kinds of reviews are written by customers who have used the particular product or service. An individual's interests, opinions and perceptions greatly influence the nature of the review. There are instances where people are biased in their opinions and automatically that has an impact on the content they contribute to the forum as review or blog posts or tweets. As the number of such people contributing content surges it has become a huge challenge to classify and organize the real problems and prospects of the product which makes the user to doubt the reliability of the content. Big companies rely on personal review of customers to improve the scope of their product and deem it to be of great importance in placing content based ads on sites that easily aid a prospective buyer.

The same applies to movie enthusiasts and voters as more and more people are using the social networking sites, online shopping and trend analysts who after reading the reviews available decide on various issues. For example placing the ad of a Kitchen Aid Mixer on a food blog not only influences purchase decisions but also goes a long way in modifying the marketing strategy.The marketing division of a company enthusiastically promotes reviewers by sending samples of product to be reviewed or sponsoring giveaways in blogs or in social networking sites like Facebook and Twitter. This has lead to the increase in the volume of data available and the need to classify the available information efficiently as these have a larger impact. The subjective nature of opinion makes a single opinion insufficient in decision making [6]. Also, the writing skills and choice of words by contributors largely depend on the language proficiency and the temperament of the writer. Online reviews that are usually the voice of the customer are written from their angle of interests and

preferences can be a combination of a positive and negative opinion which may not help in deciding whether it is a positive or a negative review. For example, consider the sentence "This restaurant's Chinese dishes are not as good as their Thai dishes". These kind of comparative opinions are different in natural language processing. When a positive word 'good' is negated like 'not as good as' a reader will also find it difficult to comprehend on how good the Thai dishes were as this decides the taste of the Chinese dishes too. When treating negation, one must be able to correctly determine what part of the meaning expressed is modified by the presence of the negation [8,2]. There are different types of opinions like regular, implicit, direct, indirect and comparative. The freedom of expression and anonymity also comes with a price. People with hidden agendas or malicious intentions to easily game the system to give people the impression that they are independent members of the public and post fake opinions to promote or to discredit target products, services, organizations, or individuals without disclosing their true intentions, or the person or organization that they are secretly working for. Such individuals are called *opinion spammers* and their activities are called *opinion spamming* [1].The rest of this paper is organized as follows: Section 2 deals with the data source. Section 3 deals with the classification methods for sentiment mining. Section 4 deals with the application of sentiment classification and section 5 deals with the evaluation and descriptions about the findings and section 6 reports the conclusions of this survey and discusses the scope for future research.

II. DATA SOURCE

Since the opinions contributed by people and companies are to be evaluated, we consider the data from blogs, review sites, web discourse and news articles.

2.1 Review

There are many user generated reviews available on the internet that aids a customer in buying a product. E-commerce sites such as www.amazon.in, www.flipkart.com and www.reviewcentre has millions of customer reviews for products, where as www.rediff.com/movies/reviews,www.indiaglitz.com and www.rottentomatoes.com has reviews for movies and www.yelp.com,www.burrp.com has restaurant reviews[23].

2.2 Web Discourse

A blog is a personal website or web page on which an individual records opinions, links to other sites, etc. on a regular basis. They are the fastest growing sections of the emerging communication systems. The simple and no-nonsense style of writing a post and uploading it on the web has made the blogging world an indispensible source of data in the case of sentiment mining [19].The micro blogging site Twitter is also flooded with opinions that are decisive in determining the election results even [3].These opinions can also be used for classifying sentiments [21]. People record the daily events in their lives and express their opinions, feelings, and emotions in an on-line journal, or blog or on Twitter [14].

2.3 News Articles

The websites like www.thesun.co.uk ,www.cnn.com and www.thehindu.com has news articles that allow users or readers to comment. This helps in recording the opinions of the people in issues that are of current relevance and importance.

III. CLASSIFICATION METHODS

In this section we review fundamental aspects of three popular supervised classifiers: Naive Bayes, Support Vector Machines and Genetic algorithm.

3.1 Naïve Bayes classification

Naive Bayes is a probabilistic learning method that assumes terms occur independently. In order to incorporate unlabelled data, the foundation Naïve Bayes was build. The task of learning of a generative model is to estimate the parameters using labeled training data only. The estimated parameters are used by the algorithm to classify new documents by calculating which class the generated the given document belongs to [4]. The naive Bayesian classifier works as follows:

1. Consider a training set of samples, each with the class labels T. There are k classes, C1, C2, . . . ,Ck. Every sample consists of an n-dimensional vector,
X = { x1, x2, . . . ,xn}, representing n measured values of the n attributes, A1,A2, . ,An, respectively.
2. The classifier will classify the given sample X such that it belongs to the class having the highest posterior probability. That is X is predicted to belong to the class Ci if and only $P(Ci|X) > P(Cj|X)$ for $1 \leq j \leq m, j \neq i$.
Thus we find the class that maximizes P(Ci |X). The maximized value of P(Ci |X) for class Ci is called the maximum posterior hypothesis.

By Bayes Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

The simplicity of the naïve bayes theorem is very useful when it comes to document classification (Hanhoon Khang Et.al (2012), Melville et al., 2009; Rui Xia, 2011; Ziqiong, 2011).The main idea is to estimate the probabilities of categories given a test document by using the joint probabilities of words and categories. The simplicity of the Naïve Bayes algorithm makes this process efficient. Hanhoon Khang Et.al (2012) has proposed an improved version of the Naïve Bayes algorithm and a unigrams + bigrams was used as the feature, the gap between the positive accuracy and the negative accuracy was narrowed to 3.6% compared to when the original Naïve Bayes was used, and that the 28.5% gap was able to be narrowed compared to when SVM was used.

3.2 Support Vector Machines

Support Vector machine is Vector space based machine-learning method aiming to find a decision boundary between two classes that is maximally far from any point in the training data (possibly discounting some points as outliers or noise). Apart from performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces .This discriminative classifier is considered the best text classification method (Rui Xia, 2011; Ziqiong, 201). M. Rushdi Saleh Et.al (2011) has applied the new research area by using Support Vector Machines (SVM) for testing different domains of data sets and using several weighting schemes. They have accomplished experiments with different features on three corpora. Two of them have already been used in several works. The SINAI Corpus has been built from Amazon.com specifically in order to prove the feasibility of the SVM for different domains [17].

3.3 Genetic Algorithm

Genetic algorithms are search heuristics that are similar to the process of biological evolution and natural selection and survival of the fittest. Genetic Algorithms (GAs) are probabilistic search methods. GAs are applied for natural selection and natural genetics in artificial intelligence to find the globally optimal solution from the set of feasible solutions [8]. The experiments with GA's start with a large set of possible extractable syntactic, semantic and discourse level feature set. The fitness function calculates the accuracy of the subjectivity classifier based on the feature set identified by natural selection through the process of crossover and mutation after each generation.

IV. APPLICATIONS

There are various tools that are readily available to evaluate the usefulness of the review. Due to the spurt of growth in the number of users online and the rampant use of the internet for decision making it becomes very important to develop an application that summarizes the reviews available. SumView is one such tool developed by Dingding Wang Et.al (2013) that summarizes the product reviews and customer opinions. It mainly focuses on summarization by delivering the majority of information contained in review documents by selecting the most representative review sentences for the extracted product feature. Different from many other systems which use benchmark datasets, SumView is a real Web-based system integrating review crawling from Amazon.com, automatic product feature extraction along with a text field where users can input their desired features, and sentence selection using the proposed feature-based weighted non-negative matrix factorization algorithm. Finally, the most representative sentences are selected to form the summary for each product feature [10].Hanhoon Kang(2012) Et.Al developed a Senti-lexicon for restaurant reviews[11]. Robert P. Schumaker Et.Al(2012) has investigated the Arizona Financial Text (AZFinText) system a financial news article prediction system, and has paired it with a sentiment analysis tool. They have found that the financial news articles have a direct impact on influencing the prices of commodities and shares.

V. EVALUATION AND DESCRIPTION

The performances of the algorithms are evaluated by many using the available metrics like precision, accuracy, F1-measure and recall. In this paper we have focused on the accuracy obtained for the algorithm. Accuracy refers to refers to the rate of correct values in the data. Table 1 depicts a clear picture regarding the recent works done in the field of sentiment mining using the techniques discussed in the algorithm.

The table also specifies the data source that was used, the feature that were selected and the mining techniques used. Regarding the efficiency of an algorithm it is very difficult to predict the best as the experiments have been carried on different datasets and using the features that the author found favorable and compatible.

Table 1 Summary of the survey

S.no	Author name & Year	Technique used	Feature selection	Data Source	Accuracy
1	Hanhoon Khang(2012)	Improved Naïve Bayes I and Naïve Bayes	Unigram, Bigram, Unigram + Bigram	Restaurant search site	81.4%
2	Xue Bai(2011)	2 stage Markov Blanket classifier	Unigram, Bigram	Movie review, Online review	92.70 %
3	Rushdi Saleh M. (2011)	SVM	Different N-gram schemes	Blogs and product reviews	91.51 %
4	Aurangazeb Khan (2011)	Naïve Bayes	Opinion terms/Expressions	Movie Review, Hotel review, airline and airport review	86.6 %
5	Kaiquan Xu(2011)	Multiclass SVM	Linguistic feature	Amazon Reviews	61%
6	Rui Xia(2011)	Naïve bayes,Maximum entropy,SVM	Unigrams, Bigrams, Dependency Grammar, Joint feature	Movie Review Multi domain dataset, Amazon reviews	NB-85.8% SVM-86.4%
7	Ziqiong(2011)	Naïve Bayes	Information Gain	Cantonese reviews	93%
8	Rudy(2009)	SVM, Hybrid	Document Frequency	Movie review, MySpace comments	89%
9	Melville(2009)	Bayesian Classification	n-grams	Blogs	91.21%
10	Songbo Tan (2008)	SVM,Centroid classifier,K-Nearest neighbourhood,Winnow	MI,IG,CHI,DI	ChnSentiCorp	SVM-90%

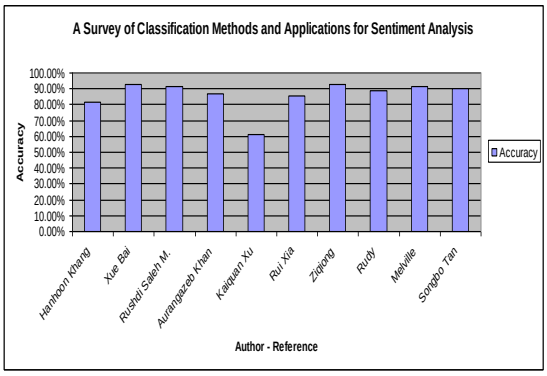

Figure:1 Classification Methods and Applications for Sentiment Analysis: A Survey

VI. CONCLUSION

Sentiment mining research is of utmost importance not only for commercial establishments but also for the common man. With the World Wide Web offering various ideas and opinions it is very important to be aware of the malicious opinions also. Based on our comprehensive literature reviews and discussions, we argue that we are actually initiating new research questions of analyzing online product reviews and other valuable online information from a domain user's point of view and exploring how such online reviews can really benefit ordinary users. In the case of product reviews there exists a visible gap between the designer's perspective and the domain user's perspective. Also that, not a single classifier can be called completely efficient as the results depend on a number of factors.

The problems in these fields of research are many. As there are numerous languages so are the syntaxes and semantics for each language. Problems like the use of negation words and the colloquial usage of certain words in a dialect affect the reviews. All these problems mentioned above can be accommodated in the future research for finding out favorable solutions.

ACKNOWLEDGEMENT

This work is supported by DST-SERB Fast track Scheme for Young Scientists by the Department of science and technology, Government of India.

REFERENCES

[1] Abbasi, A., Chen, H., Thoms, S., & Fu, T. (2008). "Affect analysis of web forums and blogs using correlation ensembles". IEEE Transactions
[2] on Knowledge and Data Engineering, 20(9), 1168–1180
[3] Ahmed Abbasi, Stephen France, Zhu Zhang and Hsinchun Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks", IEEE Transactions on Knowledge and
[4] Data Engineering, Vol. 23, No. 3, pp. 447-462, 2011.
[5] Alexander Pak and Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10, Valletta, Malta, European Language Resources Association ELRA, (May 2010).
[6] 4.Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, Opinion mining and Analysis :A survey. International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, June 2013.
[7] .Aurangzeb Khan, Baharum Baharudin and Khairullah Khan, 2011."Sentiment Classification Using Sentence-level Lexical Based Semantic Orientation of Online Reviews". Trends in Applied Sciences Research, 6: 1141-1157.
[8] .Bing Liu, Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers, May 2012.
[9] Cortes, Corinna; and Vapnik, Vladimir N.; "Support-Vector Networks", Machine Learning, 20,1995.http://www.springerlink.com/content/k238jx04hm87j80g/
[10] S Chandrakala and C Sindhu , "Opinion Mining and sentiment classification a survey" 2012 ISSN: 2229-6956(Online).ICTACT journal on soft computing.
[11] Das, Amitava, Sivaji Bandyopadhyay, and Björn Gambäck. "Sentiment analysis: what is the end user's requirement?." In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, p. 35. ACM, 2012.
[12] .Dingding Wang, Shenghuo Zhu , Tao Li," SumView: A Web-based engine for summarizing product reviews and customer opinions". Expert Systems with Applications 40 (2013) 27–33.
[13] Hanhoon Kang, Seong Joon Yoo , Dongil Han," Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews". Expert Systems with Applications 39 (2012) 6000–6010.
[14] Introduction to Information Retrieval by Hinrich Schutze (course 2010) and chapter 15 of the book [MRS08], all available at
[15] www.informationretrieval.org
[16] Lim, Ee-Peng, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady W.Lauw. "Detecting Product Review Spammers using Rating Behaviors". In Proceedings of ACM International Conference on Information andKnowledge Management (CIKM-2010). 2010.
[17] Melville, Wojciech Gryc, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification", KDD˝09, June 28–July 1, 2009, Paris, France.Copyright 2009 ACM 978-1-60558-495-9/09/06.
[18] Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog? Communications of the ACM, 47(12), 41–46
[19] Robert P. Schumaker , Yulei Zhang , Chun-Neng Huang , Hsinchun Chen," Evaluating sentiment in financial news articles". Decision Support Systems 53 (2012) 458–464.
[20] Rui Xia,Chengqing Zong, Shoushan Li, "Ensemble of feature sets and classification algorithms for sentiment classification", Information Sciences 181 (2011) 1138–1152.
[21] M. Rushdi Saleh, M.T. Martín-Valdivia, A. Montejo-Ráez, L.A. Ureña-López,"Experiments with SVM to classify opinions in different domains" Expert Systems with Applications 38 (2011) 14799–14804.
[22] Rudy Prabowo, Mike Thelwall, "Sentiment analysis: A combined approach .", Journal of Informetrics 3 (2009) 143–157.
[23] Singh and Vivek Kumar, "A clustering and opinion mining approach to socio-political analysis of the blogosphere". Computational Intelligence and Computing Research (ICCIC), 2010 IEEE
[24] International Conference.
[25] Songbo Tan , Jin Zhang, "An empirical study of sentiment analysis for chinese documents ", Expert Systems with Applications 34 (2008) 2622–2629.
[26] G.Vinodhini and Rm.Chandrasekaran "Sentiment Analysis and Opinion Mining: A Survey", Volume 2, Issue 6, June 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.
[27] 22.Xue Bai."Predicting consumer sentiments from online text". Decision Support Systems, 50, (2011), 732–742.
[28] Ziqiong Zhang, Qiang Ye, Zili Zhang, Yijun Li, "Sentiment classification of Internet restaurant reviews written in Cantonese", Expert Systems with Applications xxx (2011) xxx–xxx.