Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 5, Issue. 2, February 2016, pg.159 – 166

Review of Sentimental Analysis Methods using Lexicon Based Approach

Rahul Rajput, Arun Kumar Solanki

School of ICT, Gautam Buddha University, India

School of ICT, Gautam Buddha University, India

rahulrajput18@gmail.com, asolanki@gbu.ac.in

Abstract--- ***In our daily life we take opinion of our friends and are influenced by them in our decision making process. Opinion is the view or judgement about something. With the advent of web 2.0, the number of Social Networking Sites (SNS) has increased manifold. With these has increased the volume of users generated content. Sentiment Analysis (SA) or Opinion Mining (OM) is the computational analysis of public emotions and attitude towards a particular subject. SA is immensely useful in social media monitoring as it allows us to gain an insight of the public opinion behind certain topics. This survey paper tackles a comprehensive overview of 'Lexicon Based Approach' and the last update in it. Many recently proposed enhancements, challenges and various SA applications are investigated and presented briefly in this survey.***

Keywords--- ***Sentiment Analysis, Opinion Mining, Lexicon Based, Semantic, Sentiment Classification***

I. INTRODUCTION

Online information is becoming progressively dynamic and the emergence of online social media and user-generated content further aggravates this experience. It is hard for a person or an organization to get the latest trends and outline the general opinions about products due to the huge diversity and size of social media, and this builds the need of automated and real time opinion extraction and mining. There are number of articles presented every year in the SA fields. The number of articles in this has increased manifold. This creates a need to have survey papers that summarize the recent research trends and directions of SA.

The data sets which are used is an important part of SA. The main sources of data are from the product reviews. The reviews given by the users gives insight into product reception and quality, which can be used to make important business related decisions. The reviews sources are mainly review sites. With the explosive growth of user generated messages, Twitter has become a social site where millions of users can exchange their opinion. SA on Twitter data has provided an economical and effective way to expose public opinion. There have been some research work focusing on assessing the relations between online public sentiment and real-life events (e.g., consumer confidence, stock market [1], polls [2]). It is reported that events in real life indeed have a significant and immediate effect on the public sentiment in Twitter. Based on such correlations, some other work [3], [4] made use of the sentiment signals in blogs and tweets to predict movie sales and elections. The pioneering work of figuring out application and challenges in the field of SA was presented by Pang and Lee [5] and Liu [6]. They mentioned the techniques used to solve each problem in SA.

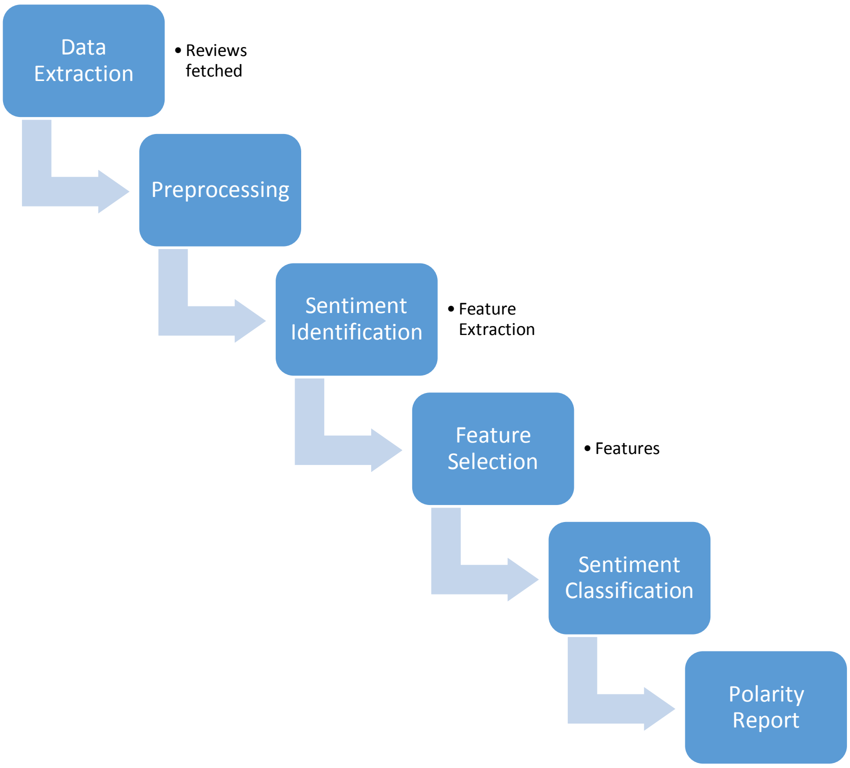


Fig. 1.

SA process can be illustrated as in Fig. 1.

II. SENTIMENTAL ANALYSIS PROCESS

Step1: Data Extraction

The data available can be fetched from the E- Commerce websites from their product review page. The data from the SNS can be extracted using Application Programming Interface (API).

For example in the case of Twitter we can extract data using Twitter API. This datum is stored in the database for further processing.

Step2: Pre-processing

Pre-processing is the process of cleaning the data readying the text for classification. Online texts contain usually lots of noise and unnecessary parts such as tags, scripts. Pre-processing the data reduces the noise in which helps to improve the performance of the classifier. Pre-processing also speeds up the classification process, thus helping in real time SA. In [7], Haddi et al. have shown that appropriate text pre-processing including data transformations and filtering can significantly improve the performance.

As instance, for a system which gives SA of twitter feeds for English tweets, we propose following pre-processing strategy:

a) *Removing Non English words*- Since we are focussing only calculating the SA of the English tweets, we must get rid of the non-English tweets.

b) *Removing Uniform Resource Locators (URLs), hashtags, references, special characters*- Cleaning the data of hashtags, references, special characters, will help reduce most of the noise. The term 'RT', which often occur in the twitter feeds should also be replaced by null.

c) *Slang word translation*- For this we take help of the internet slang dictionary and replace the slang words into their meaningful format.

d) *Removing extra letters from words*- Words which have same letter more than two times and not present in the lexicon are reduced to the word with the repeating letter occurring just once. For example, the exaggerated word "Happyyyyyy" is reduced to "Happy".

e) Stemming - Stemming is done using Natural Language Tool Kit (NLTK). Stemmer gives the stem word. For example words such as 'waiting', 'waits', 'waited' are replaced with word 'wait'.

Step3: Sentiment Identification

We aim to find out the opinionative words or phrases that best describes the context which we are dealing with. Sentiment Word Identification (SWI) is a basic technique in many SA applications. Most existing researches exploit seed words, and lead to low robustness. However Yu H et al. in [8] have proposed to identify sentiment words from the corpus without seed words.

Step4: Feature Selection

Feature selection is mostly integrated in Machine Learning (ML) algorithms like SVM, Neural Networks (NN), k-Nearest Neighbours (KNN), etc. as the very first step. As pointed out in [9], the main goal of the feature selection is to decrease the dimensionality of the feature space. Smaller feature space cuts down the computational cost. As a second objective, feature selection will also reduce the over-fitting of the learning scheme to the training data. During this process, it is also important to find a good trade-off between the richness of features and the computational constraints involved when solving the categorization task.

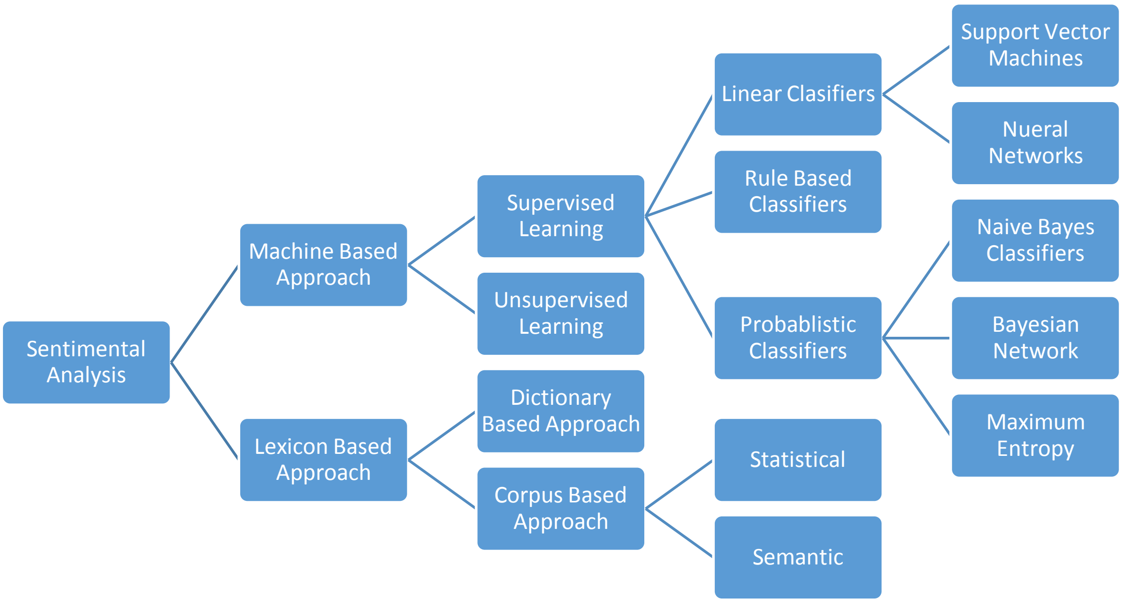
Step5: Sentiment Classification

Sentiment Classification (SC) techniques can be divided into parts namely, ML approach and lexicon based approach. ML methods are based on training an algorithm, mostly classification on a set of selected features for a specific mission and then test on another set whether it is able to detect the right features and give the right classification. In the classic work [10] Pang et al. have contributed to SC using ML Techniques. ML employs in ML algorithms using linguistic attributes while Lexicon Based approach takes cue from 'sentiment lexicon'. The ML approach can be further divided into parts namely supervised and unsupervised learning methods. The supervised learning methods make use of a large number of labeled training documents. In the case where it is difficult to find the labeled training documents, the unsupervised methods are used.

Sentiment extraction involves spotting sentiment words within a particular sentence. This is typically achieved using a dictionary of sentiment terms and their semantic orientations. Dictionary-based approach has some disadvantages associated with them. For example, the sentiment word 'low' in the context of "calories" might have a positive polarity, whereas “low” in the context of “video resolution” is of negative polarity. Taking another example, “go read the book” most likely indicates positive sentiment for book reviews, but negative sentiment for movie reviews.

III. SENTIMENT CLASSIFICATION TECHNIQUES

The SC techniques, shown in Fig. 2:



LEXICON-BASED APPROACH

The lexicon-based approach involves calculating orientation for a document from the semantic orientation of words or phrases in the document [11]. Dictionaries for lexicon-based approaches can be created manually, as authors describe in this article [12] or automatically, using seed words to expand the list of words. Much of the lexicon-based research has focused on using adjectives as indicators of the semantic orientation of text. First, a list of adjectives and corresponding Sentiment Orientation (SO) values is compiled into a dictionary. According to previous study, adjectives are good indicators of SO [13]. Then, for any given text, all adjectives are extracted and annotated with their SO value, using the dictionary scores. The SO scores are in turn aggregated into a single score for the text. However, although an isolated adjective may indicate subjectivity, there may be insufficient context to determine semantic orientation. As pointed out by Turney P. [11], the adjective ‘brutal’, ‘insane’ may have negative orientation in a pet review, in a phrase such as “brutal and insane breed of dog”, but it could have largely a positive orientation in a movie review, in a phrase like “brutal and insane action sequence”. Therefore the algorithm extracts two consecutive words. The first member is an adverb or an adjective while the second word provides the context.

A) Dictionary-based approach

Dictionary-Based approach involves using a dictionary which contains synonyms and antonyms of a word. Thus, a simple technique in this approach is to use a few seed sentiment words to bootstrap based on the synonym and antonym structure of a dictionary. Specifically, this method works as follows: A small set of sentiment words (seeds) with known positive or negative orientations is first collected manually, which is very easy. The algorithm then grows this set by searching in any online available dictionary for their synonyms and antonyms. The seed list will be added with the new found words. The process iteratively keeps on adding the words until no more new words are found. Manual inspection can be used to clean up the list at last.

B) Corpus Based Approach

The Corpus-based approach helps to solve the problem of finding opinion words with context specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus. There are two methods in the corpus based approach:

B.1) Statistical Approach

If the word appears intermittently amid positive texts, then its polarity is positive. If it appears frequently among negative texts, then its polarity can be considered as negative. If it has equal frequencies, then it can be considered as neutral word. Seed opinion words can be found using statistical techniques. Most state of the art methods are based on the observation that similar opinion words mostly appear together in a corpus. Thus, if two words appear together frequently within the same context, then there is high probability that they have same polarity. Therefore, the polarity of an unknown word can be determined by calculating the relative frequency of co-occurrence with another word. This could be done using Pointwise Mutual Information (PMI) as in example suggested by [11], SO of a given phrase is calculated by comparing its similarity to a positive word (“Awesome”) And its similarity with negative word (“Awful”). More explicitly, a phrase is given a numerical rating by taking the mutual information between the given phrase and the positive reference word “Awesome” and subtracting the mutual information between the given phrase and the negative reference word “Awful”. Using part-of-speech (POS) patterns, this technique then classifies the text by extracting the bigrams. PMI is then calculated by using the polarity score for each bigram.

B.2) Semantic approach

This principle assigns similar sentiment values to semantically close words. These Semantically close words can be obtained by getting the list of sentiment words and iteratively expanding the initial set with synonyms and antonyms and then determining the sentiment polarity for an unknown word by the relative count of positive and negative synonyms of this word [14].

IV. REVIEWED WORK

References	Methods Used	Dataset Used	Scope
15	Lexicon-based, semantic	Experienceproject.com	Stories
16	Lexicon-based, semantic	IMDB	Movie Reviews
17	Statistical (MM), semantic	Amzaon.com	Product Reviews
18	Statistical	Amazon.com	Book reviews
19	Corpus-based	Live journal blogs	Blogs
20	Lexicon-Based, SVM	Dutch wordnet	Lexicons
21	Corpus Based	Twitter	Tweet Reviews
22	Semantic	TREC 2006, TREC 2007, and TREC 2008	Blog Posts
23	Semantic	N/A	Fast Food Reviews

V. CHALLENGES IN SENTIMENTAL ANALYSIS

The main challenges that are faced by OM and SA are the following:

A) Detection of spam and fake reviews

The web contains both authentic and spam contents. For effective SC, this spam content should be eliminated before processing. This can be done by identifying duplicates, by detecting outliers and by considering reputation of reviewer.

B) Domain- independence:

The biggest challenge faced by OM and SA is the domain
Dependent nature of sentiment words. One features set may give very good performance in one domain, at the same time it perform very poor in some other domain.

C) Temporal Relations

The time of reviews may be important for SA. The reviewer may think that iPhone 5s was good in 2013, but now he may have negative opinion in 2015 because of new iPhone 6s. So assessing this kind of opinions that are changed with time may improve the performance of the SA system. This helps us to observe if a certain product gets improved with time, or people change their opinion about a product. Tan S et al. has proposed a novel work in this field [24].

D) Sarcastic sentences

Text may have Sarcastic and ironic sentences. For example, “Movie was so awesome that I had to sleep through it to forget it.” In such case, positive words can have negative sense of meaning. Sarcastic or ironic sentences can be hard to identify which can lead to erroneous opinion mining. Kreuz et al. [25] studied the role that different lexical factors play, such as interjections (e.g.“gee” or “gosh”) and punctuation symbols (e.g., ‘?’) in recognizing sarcasm in narratives. Lukin et al. in [26] explored the potential of a bootstrapping method for sarcasm classification in social dialogue to learn lexical N-gram cues associated with sarcasm (e.g., “oh really”, “I get it”, “no way”, etc.) as well as lexico-syntactic patterns.

E) Knowledge Base

Knowledge about worlds’ facts, events, people are often required to correctly classify the text. Consider the following example [27],“Casablanca and a lunch comprising of rice and fish: a good Sunday “The system without world knowledge classifies above sentence as positive due to the word “good”, but it is an objective sentence because Casablanca is the name of the famous movie.

VI. APPLICATIONS OF SENTIMENTAL ANALYSIS

SA can be used in diverse fields for various purposes. This section discusses some of the
Common ones.

A) Online Commerce

The most general use of SA is in e-commerce enterprise. Websites allow their users to submit their experience about shopping and reviewing their thoughts, opinions and their take on product qualities. They provide summary for the product and different features of the product by assigning ratings or scores. For example, htttp://www.flipkart.com is an online shopping website where in users rate the products they have bought critically.

B) Voice of Customers (VOC) and Voice of Market (VOM)

VOC is a market research technique to describe the in-depth process of capturing a customer's intentions, desires, antipathies and expectations while VOM means that you would be surveying not only your own customers but those of key competitors as well. Detection of such information as early as possible helps in direct and target key marketing campaigns.

C) Brand Reputation Management(BRM)

BRM helps in finding how public perception of a certain brand changes positively or negatively. The variation after an event can be analysed using SA. A novel work in interpreting the change in sentiment variation has been done by Tan S et al. [24]

D) Recommendation Systems

By classifying the people’s opinion into positive and negative, the system can say
Which one should get recommended and which one should not get recommended [5]

E) Policy Making

Using SA, policy makers can go through the public sentiment towards a policy, and use it to make the policies which are in demand by the public at large. [28].

REFERENCES

[1] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” J. Computer Science., vol. 2, no. 1, pp. 1–8, Mar. 2011.

[2] B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, “From tweets to polls: Linking text sentiment to public opinion time series,” in Proc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.

[3] G. Mishne and N. Glance, “Predicting movie sales from blogger sentiment,” in Proc. AAAI-CAAW, Stanford, CA, USA, 2006.

[4] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment,” in Proc. 4th Int. AAAI Conf. Weblogs Social Media, Washington, DC, USA, 2010.

[5] Pang B, Lee L, “Opinion mining and sentiment analysis” in Found Trends Inform Retriev, 2 (2008), pp. 1–135

[6] Liu B, “Sentiment analysis and opinion mining” in Synth Lect Human Lang Technol (2012)

[7] Haddi E, Liu Xi, Shi Y, “The Role of Text Pre-processing in Sentiment Analysis” in ITQM2013, Science Direct, 2013.

[8] Yu H, Deng Z, Li S, “Identifying Sentiment Words Using an Optimization-based Model without Seed Words” in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 855–859,Sofia, Bulgaria, August 4-9 2013, Association for Computational Linguistics.

[9] Kummer O, Savoy J,” Feature Selection in Sentiment Analysis”, 2000.

[10] Pang B , Lillian L, “Thumbs up? Sentiment Classification using Machine Learning Techniques” in Proceedings of EMNLP 2002, pp. 79–86.
[11] Turney P, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews” in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002.
[12] Goyal A. and Daume III H, “Generating Semantic Orientation Lexicon using Large Data and Thesaurus”.
[13] Hatzivassiloglou V, McKeown K, ” Predicting the Semantic Orientation of Adjectives”.
[15] Kim S, Hovy E. Determining the sentiment of opinions. In: Proceedings of interntional conference on Computational Linguistics (COLING’04); 2004.
[14] Neviarouskaya Alena, Prendinger Helmut, Ishizuka Mitsuru. Recognition of Affect, Judgment, and Appreciation in Text. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010), Beijing; 2010. p. 806–14.
[16] Heerschop B, Goossen F, Hogenboom A, Frasincar F, Kaymak U, de Jong F. Polarity Analysis of Texts using Discourse Structure. In: Presented at the 20th ACM Conference on Information and Knowledge Management (CIKM’11); 2011.
[17] Zirn C, Niepert M, Stuckenschmidt H, Strube M. Fine-grained sentiment analysis with structural features. In: Presented at the 5th International Joint Conference on Natural Language Processing (IJCNLP’11); 2011.
[18] Nan Hu, Indranil Bose, Noi Sian Koh, Ling Liu “Manipulation of online reviews: an analysis of ratings, readability, and sentiments” in Decis Support Syst, 52 (2012), pp. 674–684
[19] Fazel Keshtkar, Diana Inkpen “ A bootstraping method for extracting paraphrases of emotion expressions from texts”, Comput Intell, vol. 0 (2012)
[20] Alexandra Balahur, Jesús M. Hermida, Andrés Montoyo in “Detecting implicit expressions of emotion in text: a comparative analysis”, Decis Support Syst, 53 (2012), pp. 742–753
[21] Alexander Pak, Patrick Paroubek, “Twitter as a Corpus for Sentimental Analysis and Opinion Mining”
[22] Chenlo J, Hogenboom A, Losada D, “Sentiment-based ranking of blog posts using rhetorical structure theory”. in 18th international conference on applications of Natural Language to Information Systems (NLDB’13); 2013.
[23] Mao-Yuan Pai, Hui-Chuan Chu, Su-Chen Wang, Yuh-Min Chen,” Electronic word of mouth analysis for service experience” in Expert Syst Appl, 40 (2013), pp. 1993–2006
[24] Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Member, IEEE, Jiajun Bu, Member, IEEE, Chun Chen, Member, IEEE, and Xiaofei He, Member, IEEE , “Interpreting the Public Sentiment Variations on Twitter” in IEEE transactions on knowledge and data engineering, vol. 26, no. 5, may 2014.
[25] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, Ruihong Huang, “Sarcasm as Contrast between a Positive Sentiment and Negative Situation” in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)
[26] Stephanie Lukin, Marilyn Walker, “Really? Well. Apparently Bootstrapping Improves the Performance of Sarcasm and Nastiness Classifiers for Online Dialogue” in Proceedings of the Workshop on Language in Social Media (LASM 2013), pages 30–40,Atlanta, Georgia, June 13 2013. C ©2013 Association for Computational Linguistics.
[27]A. Joshi, A.R. Balamurali, P. Bhattacharyya, and R.Mohanty, “ C-feel-it: a sentiment analyzer for microblogs,” In Proceedings of ACL: Systems Demonstrations, HLT’11, 2011, pp. 127–132.
[28] Haseena Rahmath P,” Opinion Mining and Sentiment Analysis-Challenges and Applications” in International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 3, Issue 5, May 2014.