IJCTA, 10(8), 2017, pp. 185-192 ISSN: 0974-5572
© International Science Press

Sentiment Analysis of Real World Big Data–A Review of General Approaches

Padmavati Dandannavar[1] **and S.R. Mangalwede**[2]

ABSTRACT

We live in a "data age". There is a flood of data being generated every minute. The last few years have witnessed a tremendous increase in the use of Online Social Networks (OSN's). People are taking to OSN's like never before to express their opinions about products, movies, social events, services etc. This data can be aggregated and analyzed to gain valuable insights and identify opinion of public on a particular topic of interest. This need to capture public opinion has resulted in the emergence of Sentiment Analysis (SA), which detects opinions given a piece of text. A number of methods and techniques for SA have been proposed and enhanced by several researchers. This paper presents a survey of the various methods that can be used for SA of text. The advantages and drawbacks of each such method are also summarized; challenges and scope are also highlighted.

Index Terms**:** Sentiment Analysis, Lexicon based methods, Machine Learning, Keyword spotting, Concept-based methods.

1. INTRODUCTION

One lazy morning, as you flip the pages of the newspaper, an image of your favorite actor endorsing the latest model of your favorite brand of mobile phone draws your attention. You decide that you HAVE to make this purchase, but not too sure if the new model is worth buying or not. You would probably want to know what other people are opining about the new phone - after all you would not want to waste your money on something that would not meet your expectation and subsequently turn out to be not worth buying ! Before the Internet arrived on the scene, you would turn to your friends and relatives for product recommendation and whose "opinion" is important for making that decision. In the era of the Internet, you will however want to find out "online reputation" of the product which will help you make this informed decision of whether to buy or not. This online reputation is derived from the opinion of others expressed as reviews, comments, ratings online and is generally considered to represent the overall opinion about the product by the online community.

Today's consumer is spoilt with choices for products. So when a purchase decision regarding a particular product has to be made, he/she would want to base this decision on what people think about the same product. Opinions and experiences of other's are very valuable in the decision making process. Easy access to the Internet coupled with the growing popularity of social sites (Twitter, Facebook etc) and websites (like Amazon.com and Eopinion.com), people are taking to the web like never before to express their opinions and reviews on different products, movies, politics, social events etc. This has ultimately led to an ever increasing amount of data. This data is unstructured [1] and is not directly processable by machines. This kind of data captures the opinion of public - about product preferences, movies, social events etc.

This large volume of data can be tapped to gain an understanding of the sentiment expressed in the text, which has opened up the research avenue called "Sentiment Analysis" (SA) - detecting sentiments expressed in messages on online social network. SA emerged way back in 1990's but became a major field from 2004

*,** KLS Gogte Institute of Technology , Belagavi–590008, *Emails: padmad@git.edu, mangalwede@git.edu*

onwards [2]. SA refers to extracting sentiment (positive/negative/neutral) from text (phrase/sentence/document) towards an entity (person/idea/sub topic).

Navigating through several websites and reading through numerous user opinions is however both time consuming and frustrating. SA, as a process, automatically identifies opinion about an entity in user-generated text. The task of SA can be applied at different levels from document level (the basic unit of information used is the whole document and it is classified either into positive or negative class) to the sentence level (a sentence is first classified as being subjective or objective and then classified into positive, negative or neutral class) and Aspect or Feature level (deals with identifying and extracting product features from the source data).

Despite several research attempts and development of different methods for SA, it is not clear which method is better. Each method has its own pros and cons. This paper attempts to present a survey of the various methods that are commonly used for SA highlighting the pros and cons of each. The rest of the paper is organized as follows: Section 2 presents a summary of the work previously carried in the field of SA. Section 3 presents the various methods that are currently used in SA. Section 4 highlights the challenges encountered in the process of SA and section 5 presents the metrics used to evaluate the performance of SA methods.

2. LITERATURE SURVEY

Over the last few years, several methods and techniques that can be used for SA have been proposed and enhanced by several researchers. Presented in Table 1, is the summary of the papers considered for this work.

Table 1
Summary of Literature Review

Ref #	*Type of paper*	*SA approaches identified/ cited*	*Additional information/ Major Finding*	*Proposed approach/ New Contribution*
1.	Survey highlighting new avenues in SA	• Keyword Spotting • Lexical Affinity • Statistical methods Concept based methods	–	Knowledge based approaches to SA
2.	Survey	• Sentiment classification. • Feature based Sentiment • Opinion Summarization	–	–
3.	Survey	• Lexicon Based - Dictionary & Corpus based • Machine Learning Based - Supervised & Unsupervised	SA task consists of extraction, pre processing, analysis & knowledge discovery as sub tasks.	–
4.	Survey & Comparison	Machine learning, lexicon-based, statistical and rule based approaches. Based on the structure of the: text: document level, sentence level or word/ feature level	Analysis of six solutions which are representative for the categories mentioned .	–
5.	Research	• Lexicon based • Machine learning based • Hybrid methods	A hybrid method for SA in Facebook	Senti Buk, a Facebook application tool developed.

(*contd...Table 1*)

Ref #	*Type of paper*	*SA approaches identified/ cited*	*Additional information/ Major Finding*	*Proposed approach/ New Contribution*
6.	Performance comparison of three versions of a lexicon-based classifier	• Lexicon based approach • Machine learning algorithms • Hybrid methods	–	Product review (web reviews) written in Brazilian Portuguese.
7.	Survey	Document Level sentiment classification i. Supervised methods ii. Un-supervised methods Sentence Level sentiment classification *Sentence lexicon construction :* i. Corpus based ii. Dictionary based Aspect Level sentiment classification	–	–
8.	Research	• Lexicon based • Machine learning based	Describes the process of dictionary creation.	Extends the Semantic Orientation CALculator (SO-CAL) to other parts of speech; introduces intensifiers and refined approach to negation.
9.	Comparison	• Lexicon based • Machine learning based	Eight popular sentiment analysis methods were compared. Parameters used for comparison were coverage & agreement.	iFeel - a free web service providing an open API. Given a piece of text this API compares results of different SA methods.
10.	Survey	• Lexicon based • Machine learning based • Hybrid techniques	Problems in SA as stated by other researcher's in their work.	–
11.	Comparison	• Methods based on machine learning • Methods based on semantic orientation	–	A data-driven approach: a TF/IDF terms selection procedure applied in order to make computation more efficient and to improve the classification results.
12.	Review	• Keyword Spotting • Lexical Affinity • Statistical methods	Levels of SA	–
13.	Survey	Based on the usage of lexicon andRequirement of training sets	Challenges encountered in SA discussed.	–
14.	Research	• Lexicon Based	–	Holistic Lexicon-based approach to overcome drawback #2 of Lexicon based approaches
15.	Survey	Different levels of SA : sentence level, document level, aspect level & user level	List of API's, lexicons, tools that can be used in SA	–

(*contd...Table 1*)

Ref #	*Type of paper*	*SA approaches identified/ cited*	*Additional information/ Major Finding*	*Proposed approach/ New Contribution*
16.	Survey	• Lexicon based • Machine learning based • Hybrid techniques	Comparison of SA techniques. Parameters used for comparison - Recall, Precision, F-measure & Accuracy.	–
17.	Research	• Formal Approaches • Machine learning Approaches	–	Logical approach with focus on combinatory categorical grammar, lexicon acquisition & annotation and semantic networks for analyzing the sentiments of the text
18.	Review	• Lexicon based approach • Machine learning based approach	API's for collecting tweets	–
19.	Research & Comparison	• Lexical approach • Machine learning approaches	eNulog - a blog visualization tool	Approach based on Support Vector Machine (SVM)

3. APPROACHES TO SENTIMENT ANALYSIS (SA)

After review of papers cited in the reference section, it is found that the task of SA can be performed using either the Lexicon/ Lexical based methods, Machine Learning based methods, Keyword Spotting or Concept-based methods.

3.1. Lexicon based methods

Data is classified based on the number of "opinion words" present in the text [3]. Words expressing either a positive/negative sentiment are called opinion words -"excellent", "brilliant", "bad", "expensive" are examples. The semantic orientation of individual words in a review is used to collectively determine the overall sentiment polarity of the review [4]. These methods thus require a corpus [dictionary of words/ sentiment lexicon], in which each word is annotated with its semantic orientation (positive/negative) [5].

Given a piece of text whose sentiment has to be determined, these methods identify the opinion words in the text. The corpus is then used to determine the polarity (strength/ semantic orientation) of each opinion word. A semantic orientation score of +1 and-1 are assigned to the positive and negative words respectively. If the total polarity score (tps) is positive (i.e., there are more positive opinion words than negative words), the text is classified as positive, else negative.

The sentiment lexicon is the basic and most important resource for sentiment classifiers, which can be created manually/automatically [6]. The sentiment lexicon consists of sentiment words and phrases [7]. Different methods that can be used to construct a sentiment lexicon are - Manual Construction, Corpus based and the Dictionary based methods.

The advantages of Lexicon based methods are:

1) The method is simple and efficient and gives reasonable results,
2) They rely only on labeled data and do not need any training dataset, and
3) The need for a corpus may seem to be a drawback. The easy availability of lexicons and their extensibility compared to training set, prove otherwise.

Summarized below are the drawbacks of Lexicon based methods:

1) Context dependent opinion words cannot be dealt with,
2) Sentences that contain multiple conflicting opinion words cannot be dealt with.
3) A single sentence might address multiple entities and the opinion for each entity can vary.
4) One major criticism raised against these methods is that the dictionaries are unreliable as they are hand ranked by humans [8],
5) It is hard to create a unique lexical-based dictionary to be used for different contexts [9],
6) Limited words coverage -may fail to recognize words that are not already in the lexicon,
7) They usually perform less accurately than machine learning approaches [5], and
8) Results could be over/under analyzed leading to a decrease in the performance if the dictionary is too exhaustive/sparse.

3.2. Machine Learning/ Learning based methods:

Here, sentiment detection is a binary classification task i.e., positive or negative. A given piece of text is classified on the basis of labeled examples. A training dataset is used to extract the relevant features and then used to train the algorithm (Naive Bayes, Maximum Entropy, SVM etc).

Supervised, Semi-supervised and Unsupervised methods are different categories of machine learning based SA methods [10].

i. Supervised - using labeled trained data,
ii. Unsupervised - without labeled data, and
iii. Semi-supervised - mixed of labeled and unlabeled data.

The advantages of Machine Learning based methods are:

1) Machine Learning method improve the accuracy significantly [3],
2) Ability to adapt and create trained model for specific purposes and contexts [9],
3) Better results can be obtained using machine learning approaches in bounded domains [5], and
4) Machine learning methods give more accuracy [11].

The drawbacks of Machine learning methods are:

1) Supervised methods require labeled inputs as training data - the larger the better,
2) If there is a change in the language use, these methods adapt very poorly,
3) Availability of labeled data and hence the low applicability of the method on new data [9],
4) Labeling data is costly and prohibitive for some tasks,
5) Machine learning methods rely on the use of a large dataset of labeled documents for training [11],
6) When a classifier that has achieved a high accuracy is used for another domain, its performance decreases [6] and,
7) Collecting and annotating a large corpus of tagged training data can be both, a challenging and expensive task.

3.3. Keyword Spotting

This method includes developing a list of affect words (keywords that relate to a certain sentiment). These are usually positive/negative adjectives an can be strong indicators of sentiment [12].

Limitations of this approach are:

1) When negation is involved, affect words are poorly recognized [12].

   For example, the sentence “today’s weather was good” will be correctly classified as having a positive sentiment where as in trying to classify the sentence “today’s weather wasn’t good”, it is likely to fail.

2) This technique relies on the presence of affect keywords that are only surface feature of the prose [1, 12].

For eq., the sentence “My wife filed for divorce today” has no affect words but conveys strong emotions. Therefore keyword spotting is ineffective.

3.4. Concept-based method

Rely on large semantic knowledge bases and use web ontology’s/semantic networks for semantic text analysis [13]. Unlike methods that use keyword(s) and word co-occurrence counts, they use implicit meaning associated with natural language concepts.

Advantages of this approach include:

1) Detection of sentiments that are expressed subtly, and
2) Multi-word expressions can be analyzed even though they don’t explicitly convey emotions but are related to concepts that do [1].

Heavy reliance on the depth and breadth of the knowledge base used is the **major drawback** of this method.

4. CHALLENGES IN SA

The availability of several techniques for SA, makes it is difficult to say which method is better over the other. To improve the overall performance and accuracy of SA methods, it is necessary to address the following issues and challenges:

1) An opinion/affect word maybe considered positive in one situation and negative in another situation,
2) The same opinion can be expressed differently by different people. They can be contradictory in their statements,
3) People may use a single sentence that combines different opinions – which is easy for a human to understand but different for a computer/machine to parse,
4) When a short piece of text lacks context it becomes difficult for even human’s to understand what someone else thought,
5) Handling negations, polysemy (a word with multiple meanings), slangs and domain generalization,
6) Identifying hidden sentiments (e.q., anger, disgust, joy) is a challenging task, and
7) Updating / Down-dating Lexicons [13]-Since a sentiment analyzer uses a lexicon, its performance depends majorly on the accuracy of the lexicon. Lexicons should thus be updated to suppress the words which are no longer used and to accommodate new words.

5. PERFORMANCE MEASURES

To assess the accuracy of a classifier, it is common practice to create a confusion matrix. A confusion matrix (shown in Table 2), is a table that describes the performance of a classification model (or “classifier”) on a set of test data for which the true values are known [20].

Table 2
Confusion Matrix

	Predicted positives	*Predicted positives*
Actual positive Instances	TP	FN
Actual Negative Instances	FP	TN

The basic terms appearing in the confusion matrix are:

i) *True positives (TP)* - the classifier predicted yes, and it is actually a yes.

ii) *True negatives (TN)* - the classifier predicted no, and it is actually a no.

iii) *False positives (FP)* - the classifier predicted yes, but it actually is a no (also known as a “Type I error”).

iv) *False negatives (FN)* - the classifier predicted no, but it actually is a yes (also known as a “Type II error”).

Accuracy, Precision, Recall and F Score are four measures/indices used to measure the performance of the sentiment classifiers. These indices are computed based on the confusion matrix.

1) *Accuracy* - Overall, how often is the classifier correct?

$$Acc = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

2) *Precision* - When a classifier predicts yes, how often is it correct? This measure shows how accurately the model makes predictions.

$$p = TP/(TP + FP) \quad (2)$$

3) *Recall* - is the portion of true positive predicted instances against all actual positive instances [6]

$$r = TP/(TP + FN) \quad (3)$$

4) *F1* - is a weighted average of the true positive rate (recall) and precision.

$$F1 = (2 * p * r)/(p + r) \quad (4)$$

6. CONCLUSION

It can thus be concluded that to perform sentiment analysis of text, a number of techniques can be used. Some techniques use a lexicon, some use training set and some use both. But the methods are domain specific. Languages that have been studied mostly are English and Chinese. Very few researches have been conducted on sentiment classification for other mode languages. Most sentiment classifiers are implemented for English Language and hence most research on SA focuses on text written in English as a result of which most of the resources developed (example, corpora and sentiment lexicon) are in English. None of the existing technique’s are generalized enough to be language independent. Thus attempts need to be made towards a generalized Sentiment Analyzer. Other major challenges include handling negation and language generalization.

REFERENCES

[1] Erik Cambria, Björn Schuller, Yunqing Xia,Catherine Havasi : “New Avenues in Opinion Mining an Sentiment Analysis”, Published by the IEEE Computer Society March/April 2013 1541-1672/13 .

[2] G. Vinodhini, RM. Chandrasekaran : “Sentiment Analysis and Opinion Mining: A Survey”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6, June 2012 ISSN: 2277 128X.

[3] Sagar Bhuta, Avit Doshi, Uehit Doshi, Meera Narvekar : “A Review of Techniques for Sentiment Analysis “ , International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT) , 978-1-4799-2900-9/14 2014 IEEE.

[4] Anais Collomb,Crina Costea, Damien Joyeux, Omar Hasan, Lionel Brunie : “A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation”.

[5] Alvaro Ortigosa , José M. Martín, Rosa M. Carro: “Sentiment analysis in Facebook and its application to e-learning”, Elsevier Journal of Computers in Human Behavior 31 (2014), Published by Elsevier Ltd.

[6] Lucas V. Avanco, Maria G. V. Nunes: “Lexicon-based Sentiment Analysis for Reviews of Products in Brazilian Portuguese”, 2014 Brazilian Conference on Intelligent Systems. 978-1-4799-5618-0/14 DOI 10.1109/BRACIS.2014.57

[7] Mohammad Sadegh Hajmohammadi, Roliana Ibrahim, Zulaiha Ali Othman: “Opinion Mining and Sentiment Analysis: A Survey”, International Journal of Computers & Technology Volume 2 No. 3, June, 2012. ISSN: 2277–3061 (online)

[8] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll :”Lexicon-Based Methods for Sentiment Analysis”, Association for Computational Linguistics 2011.

[9] Pollyanna Gonçalves, Matheus Araújo : “Comparing and Combining Sentiment Analysis Methods”, COSN’13, October 07–08, 2013, ACM 978-1-4503-2084-9/13/10 http://dx.doi.org/10.1145/2512938.2512951.

[10] Zohreh Madhoushi, Zohreh Madhoushi, Suhaila Zainudin: “Sentiment Analysis Techniques in Recent Works” Science and Information Conference 2015 July 28-30, 2015 - London, UK.

[11] Valentina Mazzonello, Salvatore Gaglio, Agnese Augello, Giovanni Pilato: “A Study on Classification Methods Applied to Sentiment Analysis “, 2013 IEEE Seventh International Conference on Semantic Computing,978-0-7695-5119-7/19 DOI 10.1109/ICSC.2013.82

[12] Shreya Banker, Rupal Patel: “A brief review of sentiment analysis methods”, International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016. DOI : 10.5121/ijist.2016.6210 89

[13] Ms Kranti Ghag and Dr. Ketan Shah: “Comparative Analysis of the Techniques for Sentiment Analysis”, ICATE 2013 Paper Identification Number-124.

[14] Xiaowen Ding, Bing Liu, Philip S. Yu : “A Holistic Lexicon-Based Approach to Opinion Mining” , WSDM’08, February 11-12, 2008, ACM 978-1-59593-927-9/08/0002

[15] Khaled Ahmed , Neamat El Tazi , Ahmad Hany Hossny : “Sentiment Analysis Over Social Networks: An Overview”, 2015 IEEE International Conference on Systems, Man, and Cybernetics, 978-1-4799-8697-2/15 IEEE DOI 10.1109

[16] Asad Bukhari, Um-e-Ghazia, Dr. Usman Qamar : “CRITICAL REVIEW OF SENTIMENT ANALYSIS TECHNIQUES”, Proceeding of the International Conference on Artificial Intelligence and Computer Science (AICS 2014), 15 - 16 September 2014, Bandung, INDONESIA. (e-ISBN 978-967-11768-3-2).

[17] Neha R. Kasture, Poonam B. Bhilare: “An Approach for sentiment analysis on social networking sites”, IEEE International Conference on Computing Communication Control and Automation 2015. 978-1-4799-6892-3/15 2015.

[18] [Aliza Sarlan, Chayanit Nadam, Shuib Basri : “ Twitter Sentiment Analysis”, IEEE International Conference on Information Technology and Multimedia (ICIMU), November 18–20, 2014, Putrajaya, Malaysia. 978-1-4799-5423-0/14.

[19] [Michelle Annett, Grzegorz Kondrak: “A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs”, S. Bergler (Ed.): Canadian AI 2008, LNAI 5032, pp. 25–35, 2008. Springer-Verlag Berlin Heidelberg 2008.

[20] http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/