A Survey on Sentiment Analysis and Opinion Mining Techniques

Amandeep Kaur
University Institute of Engineering and Technology, Chandigarh, India
Email: amandeepk.cql@gmail.com

Vishal Gupta
University Institute of Engineering and Technology, Chandigarh, India
Email: vishal@pu.ac.in

***Abstract*—Sentiment Analysis (SA), an application of Natural Language processing (NLP), has been witnessed a blooming interest over the past decade. It is also known as opinion mining, mood extraction and emotion analysis. The basic in opinion mining is classifying the polarity of text in terms of positive (good), negative (bad) or neutral (surprise). Mood Extraction automates the decision making performed by human. It is the important aspect for capturing public opinion about product preferences, marketing campaigns, political movements, social events and company strategies. In addition to sentiment analysis for English and other European languages, this task is applied on various Indian languages like Bengali, Hindi, Telugu and Malayalam. This paper describes the survey on main approaches for performing sentiment extraction.**

***Index Terms*— Natural Languages processing, Sentiment Analysis, Indian languages.**

I. Introduction

"*Sentiment analysis or opinion mining refers to the application of natural language processing, computational linguistics and text analytics to identify and extract subjective information in source materials*"(Source:Wikipedia). Opinion mining/sentiment analysis is a multidisciplinary and multifaceted Artificial intelligence problem. Its aim is to minimize the gap between human and computer. Thus, it is collection of human intelligence and electronic intelligence for mining the text and classifying user sentiments, likes, dislikes and wishes. The user generated content is available in various forms such as web logs, reviews, news, discussion forums. Web 2.0 & 3.0 has provided a platform to share the feelings and views about the products and services. The basic of this problem can be better explained using the following review by a user about a car:

"*I bought a car a few days ago. It had a comfortable suspension set-up but it did not provide stable and safe feel. It delivers good fuel economy but does feel lethargic engine. It has better quality interiors which look sporty*."
The above review is analyzed for opinion mining and extracted views are visualized in Fig 1(a). Similarly comparison of two or more can also be evaluated as represented in Fig 1(b).

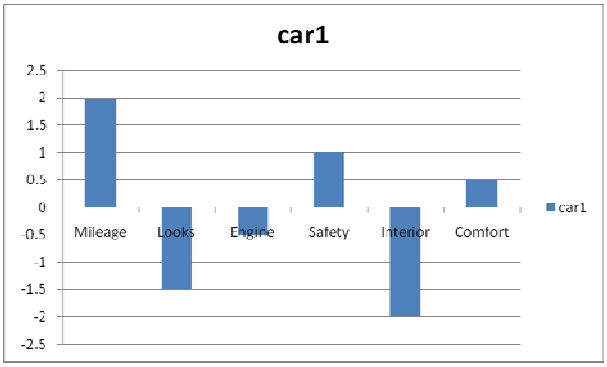

Figure 1(a). Visualization of summary of opinions on a car

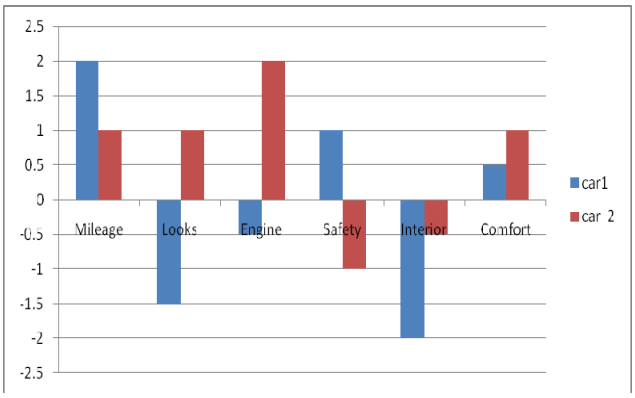

Figure 1(b). Visual comparison of two cars.

Social network revolution plays a crucial role in gathering information containing public opinion. To obtain subjective and factual information from this information, public opinions are extracted. Thus, it is the process to predict hidden information about user's intensions, likeliness and taste. These social networking sites generate enormous data up to terabytes per week. The statistics mentioned in Table I shows the rate at which amount of user generated data is increasing:

doi:10.4304/jetwi.5.4.367-371

TABLE I.
Statistics of user generated data

Facebook	Twitter	Google
850 million users 250 million photos 2.7 billion likes	465 million accounts 175 million tweets	90 million users 675000 users per day

Popular approaches used for sentiment analysis are: Popular approaches used for sentiment analysis are:

- **Subjective lexicon** – is a list of words where each word is assigned a score that indicates nature of word in terms of positive, negative or objective.
- **Using N-Gram modeling-** for given training data , we make a N-Gram model(uni-gram, bi-gram, tri-gram or combination of these)for classification.
- **Machine learning –** perform the supervised or semi- supervised learning by extracting the features from the text and learn the model.

II. Process of Sentiment Analysis for Text

The process of sentiment analysis is divided into five steps [11]: Process of Sentiment Analysis for Text (Lexicon Generation), Subjectivity Detection, Sentiment polarity Detection, Sentiment Structurization, Sentiment Summarization-Visualization-Tracking.

A. Process of Sentiment Analysis for Text (Lexicon Generation)

In this phase, sentiment lexicon is created to acquire the knowledge about sentiments. According to previous studies, prior polarity should be attached at each lexicon level. To develop SentiWordNet(s), Manual and Automated processes have been attempted for multiple languages.

Related Work: (Stone, 1966) Philip Stones developed General Inquirer system, was the first milestone for extracting textual sentiment. It was based on the manual database containing set of positive or negative orientations and the input words are compared with database to identify their class such as positive, negative, feel, pleasure [5].

(Brill, 1994) Brill Tagger depicted the semantic orientation for verbs, adverb, noun and adjective. After extracting these phrases, PMI algorithm (Turney, 2002) is applied to identify their semantic polarity [31][7].

(Hatzivassiloglou et al., 1997) Hatzivassiloglou was the first to develop empirical method of building sentiment lexicon for adjectives. The key point is based on the nature of conjunctive joining the adjectives. A log-linear regression model is provided with 82% accuracy [8].

(Turney, 2002) For the classification of positive and negative opinion, Peter Turney proposed the idea of Thumbs Up and Thumbs Down. For better problem formalization, there was the necessity of an automated system, which could be employed for electronic documents. For consecutive words and their polarity, Turney came up with an algorithm to extract Point wise Mutual Information(PMI).Experiments were conducted on movie review corpus and polarity is referred to as "thumbs up" for positive and "thumbs down" for negative [7].

(Pang et al.,2002) Pang build sentiment lexicon for movie reviews to indicate positive and negative opinion. This system motivated the other machine learning approaches like Support Vector Machine, Maximum Entropy and Naive Bayes[10].

(Kamps et al., 2004) Kamps, Marx, Mikken and Rijke tried to identify subjectivity of adjectives in Word Net. In this research, they classified adjectives into four major classes and used base words (to measure relative distance) depending on the class. For class *Feeling* their base words were “happy” and “sad”, for class *Competition* their base words were “pass” and “fail”, etc. Based on this idea, they gathered a total of 1608 words in all four classes with average accuracy of 67.18% for English [24].

(Gamon et al., 2005) proposed similar method as by (Turney, 2002).Machine Learning based technique is used with input of some seed words. This classifier is based on assumption that the words with same polarity co-occur in one sentence but words with different polarity cannot [11].

(Read, 2005) have stated three different problems in the area of sentiment classification: Time, Domain and Topic dependency of sentiment orientation. It has been experimented that associative polarity score varies with time [12].

(Denecke, 2009) introduced uses of SentiWordNet in terms of prior polarity scores. The author proposed two methods: rule-based and machine learning based. Accuracy of rule-based is 74% which is less than 82% accuracy of machine learning based. Finally, it is concluded that there need more sophisticated techniques of NLP for better accuracy [13].

(Mohammad et al., 2009) proposed a technique to increase the scope of sentiment lexicon. It includes the identification of individual words as well as multi-word expressions with the support of a thesaurus and a list of affixes. The technique can be implemented by two methods: antonymy generation and Thesaurus based. Hand-crafted rules are used for antonymy generation. Thesaurus method is based on the seed word list which means if a paragraph has more negative seed words than the positive ones, then paragraph is marked as negative [14].

(Mohammad and Turney, 2010) developed Amazon Mechanical Turk, an online service by Amazon, to gain human annotation of emotion lexicon. But there was the need of high quality annotations. Various validations are provided so that erroneous and random annotations are rejected, discouraged and re-annotated. Its output provides 2081 tagged words with an average tagging of 4.75 tags per word [10].

B. Subjectivity Detection

Sentiment analysis categorizes the text at the level of subjective and objective nature. Subjectivity means the

text contains opinion and objectivity means text contains no opinion but contains some fact. In precise form, Subjectivity can be explained as the Topical Relevant Opinionated Sentiment [9].Genetic Algorithm (Das, 2011) achieved a good success for the subjectivity detection for Multiple Objective Optimization [27].

Some example-

1. Subjective- *The car is comfortable.* (This sentence expresses the feeling which is an opinion; hence it is of subjective nature)

2. Objective- *Maruti launched a new car*. (This sentence contains a fact).

Related Work: (Wiebe, 2000) defined the concept of subjectivity in an information retrival perspective which explains the two genres subjective and objective [9].

(Aue and Gamon, 2005) told that subjectivity identification is a context dependent and domain dependent problem which replaces the earlier myth of using sentiwordnet or subjectivity word list etc. as prior knowledge database [17].

(Das and Bandyopadhyay, 2009) explained the techniques for subjectivity based on Rule-based, Machine learning and Hybrid phenomenon [2].

The idea of collecting subjectivity clues helped in the subjectivity detection. This collection includes entries of adjectives (Hatzivassiloglou and Mackeown, 1997) and verbs(Wiebe,2000) and n-grams(Dave et.al, 2003) [8][9][18].

The detail of sentiment analysis and subjectivity detection is given by Wiebe in 1990 [16].

Methods of identification of polarity are explained in (Aue and Gamon,2005) [17].

Some algorithms like Support Vector Machine (SVM), Conditional Random Field(CRF) (Zhao et. al,2008) have been used for clustering of opinions of same type[6].

C. Sentiment Polarity Detection

The sentiment polarity detection means classifying the sentiments into semantic classes(Turney et.al.,2002) such as positive, negative or neutral or other emotion classes like anger, sad, happy, surprise [7].

SentiWordNet is most popular to be used as polarity lexicon. Another technique used for polarity detection is Network Overlap Technique [27]. In this, contextual prior polarity is assigned to each sentiment word.

Related Work: Since the last few years, Tweet Feel http://www.tweetfeel.com) and Twitter Sentiment Analysis Tool(http://twittersentiment.appspot.com/) are available. To satisfy the end users, level of research should be raised [19].

(Cambria et al., 2011) developed a new paradigm known as Sentic Computing. This research is based on a common sense and emotion representation. It have been used for short texts to infer emotional states over the web [20].

Concept Net, a semantic network was introduced with approx 10000 concepts and more than 72000 features extracted from Open mind corpus. In the sentic computing, four dimensions are taken as basis to classify the affective states: Sensitivity, Attention, Pleasantness and Aptitude.

D. Sentiment Structurization

Sentiment Analysis explained till now is not sufficient to satisfy the needs of end user, because the latter is not interested in binary output in terms of positive or negative but interested in aspectual sentiment classification. Aspectual can be explained as relative information. For example, a social worker may be interested to know the change in the society before and after implementation of his scheme. So, a sentiment analysis system should be understand and identify the aspectual sentiments present in the text.

For this problem, sentiment structurization technique has been proposed by Das(Das 2010).This technique is based on 5W (Why, Where, When, What, Who).The drawback of 5Ws is that it may lead to label bias problem. To solve this Problem Maximum Entropy Model (MEMM) was introduced.

Related Work: (Bethard et. al., 2006) have introduced the automatic identification of opinions from question answering.

(Bloom et al., 2007) describes Appraisal Theory (Martin and white ,2005).The system classifies the opinions into three types: affect, appreciation or judgment.

(Yi et.al. 2006) introduced e sentiment analyzer for online text documents.

(Zhou et. al., 2006) have introduced the architecture for blogosphere to get the summarized text.

E. Sentiment Summarization-Visualization-Tracking

One of the main needs of end users is the aggregation of data. After the Literature survey, following summarization attempts are found:

- Polarity wise (Hu, 2004),(Yi and Niblack, 2005),(Das and Chen, 2007)
- Topic wise(Yi et al.,2003),(Pang and Lee, 2004) ,(Zhou ,2006)

Visualization and Tracking is the last phase of sentiment analysis which is most important to satisfy the needs of end users. In this phase, visual sentiments are generated which are further tracked with polarity wise graph according to some dimension or combination of dimensions. The final graph for tracking is created with a timeline.

III. Sentiment Analysis for Indian Languages

There is comparatively less research has been done for Indian languages.

(Das and Bandyopadhyay, 2010) suggested a computational technique for developing SentiWordNet(Bengali) using English-Bengali bilingual dictionary and English Sentiment Lexicons [21].

(Das and Bandyopadhyay, 2010)- The author introduced four approaches to predict the polarity of a word. In the First strategy; an interactive game is provided which identify the polarity of the words. In the Second strategy, a bi-lingual dictionary is developed for

English and Indian Languages. In the third strategy, word net expansion is done using antonym and synonym relations. In the fourth approach, a pre-annotated corpus is used for learning [1].

(Das and Bandyopadhyay, 2010)- developed the method for tagging using the Bengali words. Classification of words is performed into six emotion classes (happy, sad, surprise, fear, disgust, anger) according to three categories of intensities (low, general and high) [22].

(Draya et.al., 2009) performed blog sentiment analysis to extract domain specific adjectives [23].

(Joshi, et.al. 2010) used two lexical resources: English-Hindi Word Net Linking and English SentiWordNet and created H-SWN(Hindi-SentiWordNet) [28].

(Kim and Hovy,2004) Kim and Hovy did the research work for Hindi Language but their work is restricted to synonyms [24].

(Narayan, et.al., 2002) Hindi Subjective Lexicon and hindi WordNet is used for the identification of semantic orientation of adjectives and adverbs [25].

(Pang, et.al. 2002) sentiment classification is done at document level using syntactic approach of N-Grams. This method is used to perform machine learning [10].

(Rao and Ravichandran, 2009) performed the classification of bi-polar nature [26].

(Turney, 2002) Turney used semantic mining for binary classification and also did research on part of speech (POS) information. It is document level and review level sentiment analysis [7].

IV. Conclusion and Future Work

A. Conclusion

Sentiment Analysis has lead to development of better products and good business management. This research area has provided more importance to the mass opinion instead of word-of-mouth.

In the conclusion, it has been proved that coverage expansion is good by using automatic processes where as prior polarity assignment is credible by using manual methods. SentiWordNet has been successfully generated for Hindi, Telugu and Bengali and global SentiWordNet has been generated for 57 languages. The 5W structure is more acceptable solution across domains. The success of Genetic Algorithm can be estimated by the fact that the system based on this algorithm has highest performance till date for Bengali and English.

For Indian languages, scarcity of resources has become the biggest issue. Research is going on for building subjective lexicon and datasets for Indian languages.

B. Future Work

As SentiWordNet has been generated for various languages, there can be further research on cross-lingual sentiment sense mapping. It is necessary to update the prior polarity scores according to various dimensions. The future research can be to develop web service API so that latest prior polarity scores can be accessed. The concept of Artificial intelligence can be used for further research that can mimic human biological mechanism. In the future, Event Tracking can also be implemented using the concept of 5W structure.

There are 22 official languages and 13 languages having more than 10 million speakers in India. Research is going on these languages but successful results are obtained in few languages such as Bengali, Hindi and Malayalam. There are many languages which are unexplored. Multilingual dictionary is available for English and 11 Indian languages (Hindi, English, Marathi, Bengali, Gujarati, Oriya, Malayalam, Urdu, Punjabi, Tamil, and Telugu).In future; subjective lexicon can be developed for the unexplored languages which does not have a word net. The basic resources like parsers, named entity recognizers, morphological analyzers, and part of speech tagger need to be improved to reach the state of accuracy.

References

[1] A. Das and S. Bandyopadhay, "SentiWordNet for Indian languages," *Asian Federation for Natural Language Processing*, China, pp. 56–63, August 2010.

[2] A. Das and S. Bandyopadhay, "Subjectivity Detection in English and Bengali: A CRF-based Approach," In Proceedings of the 7th International Conference on Natural Language Processing, Macmillan 2009.

[3] H. Tang, S. Tan and X. Cheng, " A survey on sentiment detection of reviews", In Proceedings of the Expert Systems with Applications 36, Elsevier Ltd., Beijing, 2009

[4] N. Mohandas, J.P. Nair and G.V, " Domain Specific Sentence Level Mood Extraction from Malayalam Text", In Proceedings of the International Conference on advances in Computing and Communications, IEEE, 2012.

[5] P.J. Stone, "The General Inquirer: A Computer Approach to Content Analysis", The MIT Press, 1966.

[6] J. Zhao, K. Liu and G. Wang, "Adding Redundant Features for CRFs- based Sentence Sentiment Classication" In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp.117-126, 2008.

[7] P. Turney, "Thumbs up or thumbs down? Semantic orientation Applied to Unsupervised Classification of Reviews" In Proceedings of the Association for Computational Linguistics, pp.417-424, Philadelphia, 2002.

[8] V. Hatzivassiloglou and K. R. McKeown, "Predicting the Semantic Orientation of Adjectives", In Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the European Chapter of the ACL, pp.174-181, Madrid, 1997.

[9] J. M. Wiebe, " Learning Subjective Adjectives from Corpora", In Proceedings of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence, pp. 735-740, Mento Park, 2000.

[10] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques" In Proceedings of the Empirical Methods on Natural Language Processing, pp. 79-86, Pennsylvania, 2002

[11] M. Gamon, A. Aue, S. Corston-Oliver and E. Ringger, "Pulse: Miming Customer Opinions from Free Text", In Proceedings of the International Symposium on Intelligent Data Analysis, pp. 121-132, 2005.

[12] J. Read, "Using Emotions to Reduce Dependency in Machine Learning Techniques for Sentiment

Classification", In Proceedings of the Student Research Workshop, pp. 43-48, Arbor, 2005.
[13] K. Denecke, "Are SentiWordNet Scores Suited For Multi-Domain Sentiment Classification", In Proceedings of the 4th International Conference on Digital Information Management, pp. 33-38, Ann Arbor, 2009.
[14] S. Mohammad, B. Dorr, and C. Dunne, " Generating High-Coverage Semantic Orientation Lexicons fom Overly Marked Words and a Thesaurus", In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing,pp. 599-608, 2009.
[15] S. Mohammad and P. Turney, "Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon" In Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion, pp. 26-34, California, 2010.
[16] J. M. Wiebe, "Recognizing Subjective Sentences: A Computational Investigation of Narrative Text", Doctoral Thesis. UMI Order Number: UMI Order No. GAX90-22203., State University of New York, 1990.
[17] M. Gamon and A. Aue, " Automatic Identification of Sentiment Vocabulary: Exploiting Low Association with Known Sentiment Terms", In Proceedings of the Workshop on Feature Engineering for Machine Learning in Natural Language Processing, pp. 57-64, Ann Arbor, 2005.
[18] K. Dave, S. Lawrence and D. M. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", In Proceedings of 12th International Conference on World Wide Web, pp. 519-528, Hungary, 2003.
[19] B. Liu, "Sentiment Analysis: A Multi- Faceted Problem", In Proceeding of IEEE Intelligent Systems, pp. 76-80, 2010.
[20] E. Cambria, A. Hussain and C. Eckl, " Taking Refuge in Your Personal Sentic Corner", In Proceeding of Workshop on Sentiment Analysis where AI meets Psychology, pp. 35-43, Thailand, 2011.
[21] A. Das and S. Bandyopadhay, "SentiWordNet for Bangla", 2010.
[22] A. Das and S. Bandyopadhay, "Labeling emotion in Bengali blog corpus a fine grained tagging at sentence level", In Proceeding of 8th Workshop on Asian Language Resources, pp. 47-55, Beijing, 2010.
[23] G. Draya, M. Planti, A. Harb, P. Poncelet, M. Roche and F. Trousset, " Opinion Mining from Blogs", In Proceeding of International Journel of Computer Information Systems and Industrial Management Applications, 2009
[24] S. M. Kim and E. Hovy, "Determining the sentiment of opinions", In Proceeding of COLING, pp. 1367-1373, 2004.
[25] D. Narayan, D. Chakrabarti, P. Pande and P. Bhattacharyya, " An experience in building the indo wordnet -a wordnet for hindi", In Proceeding of First International Conference on Global WordNet, 2002.
[26] D. Rao and D. Ravichandan, " Semi-supervised polarity lexicon induction", In Proceeding of 12 conference of the European Chapter of the Association for Computational Linguistics, pp. 675-682, USA, 2009.
[27] A. Das, "Opinion Extraction and Summarization from Text Documents in Bengali", Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., Jadavpur University, 2011.
[28] A. Joshi, A. R. Balamurali and P. Bhattacharyya, "A Fall-back Strategy for Sentiment Analysis in Hindi: a Case Study", In Proceedings of the 8th ICON, 2010.
[29] Amitava Das, http://www.amitavadas.com/
[30] Janyce M. Wiebe,"Learning Subjective Adjectives from Corpora",http://www.cs.columbia.edu/~vh/courses/LexicalSemantics/Orientation/wiebe-aaai2000.pdf
[31] Brill Tagger, http://www.ling.gu.se/~lager/mogul/brill-tagger/index.html.