Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
Article DOI: 10.21474/IJAR01/1521
DOI URL: http://dx.doi.org/10.21474/IJAR01/1521
ISSN NO. 2320-5407

RESEARCH ARTICLE

SENTIMENT ANALYSIS: A GENERALISTIC REVIEW.

Karan Dharni.
Department of Computer Science and Engineering, Jaypee University of Engineering and Technology Guna, 473226, India.

Manuscript Info

Manuscript History

Received: 16 July 2016
Final Accepted: 19 August 2016
Published: September 2016

Key words:-
Opinion Mining, Sentiment, Text Classification, Knowledge Discovery

Abstract

In today's world, Internet users play a key role in development of data and information on the world wide web. As a result, detection of emotions and opinions in text has become an intensely researched field recently. And sentiment analysis is the branch of Natural Language Processing with the above mentioned objectives. Applications of this stream of knowledge discovery are vast and varied. In this paper, we have sought to provide a concise summary and a brief survey of the prominent research done in the field of Sentiment Analysis.

Introduction:-

The term 'sentiment' is defined as a thought, view or attitude premised on emotions. "Sentiment Analysis is a Natural Language Processing (NLP) task that deals with finding orientation of opinion in a piece of text with respect to a topic" [1].Sentiment analysis is an interdisciplinary field that combines the knowledge and endeavors of artificial intelligence, NLP, data mining and knowledge discovery. The primary task of sentiment analysis is to determine and identify the polarity of a piece of text as negative, positive or neutral. The year 2001 seems to mark the inception of widespread awareness and research initiatives in this field.[2]

In today's world, Internet users are no longer passive consumers of information; and instead play a key role in development of data and information on the world wide web in the form of blogs, reviews, surveys, comments etc. As a result, detection of emotions and opinions in text has become a hotly researched field in recent years. Sentiment Analysis has found application spanning a wide gamut of fields; including, though not limited to businesses, marketing, politics and even humanities. Recently, sentiment analysis techniques has been applied to predict election results [3] and summarize opinions regarding sensitive social issues such as abortion [4].

From a technical point of view, Bag Of Words (BOW) and Feature Based Sentiment (FBS) are the two approaches followed in the field of sentiment analysis [5].

In BOW approach, the whole text is seen as an collection of words, disregarding semantic relationships between words. While on the other hand, the FBS approach has emerged to analyze opinions and sentiments towards products and their features [4].

Furthermore, it is widely believed and accepted that sentiment analysis is domain specific, so the classification of a certain part of text as positive or negative, depends on the domain being considered.

Corresponding Author:- Karan Dharni.
Address:- Department of Computer Science and Engineering, Jaypee University of Engineering and Technology Guna, 473226, India.

The rest of the paper is organized as follows. In Section 2 we have discussed the methodology generally used in sentiment analysis related research works. Section 3 gives an overview of related work and provides a tabular summary of existing literature. While Section 4 and 5 discuss the future trends in sentiment analysis and the conclusions respectively.

Methodology:-

Over the past decade there has been extensive research in the stream of sentiment analysis and opinion mining. The general methodology used by researchers broadly includes, classification of text, followed by its analysis. The classification and analysis methods used are discussed below.

Classification Techniques:-

Three types of approaches for sentiment classification of texts exist, namely-

1. Supervised machine learning: It includes using the Naive Bayes, Maximum Entropy and Support Vector Machines (SVM) .
2. Lexicon- based: It includes using an unsupervised semantic orientation scheme or using publically available libraries such as SentiWordNet for labeling text as positive negative or neutral.
3. Hybrid Approach: This approach attempts to combine both the above mentioned techniques into a single framework.

Abstractly, Naive Bayes is a conditional probability model. Using Bayes Theorem, conditional probability can be expressed as:

$$p(C_k/X) = (P(C_k) * P(x/C_k)) / p(X)$$

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

To formally define a hyperplane, the following notation is used:

$$f(x) = \beta_0 + \beta^T x,$$

where β is known as the *weight vector* and β_0 as the *bias*.

Analysis Techniques:-

In the process of determining the innate sentiment within a piece of text, the following three factors are used-

1. Subjectivity: Subjective analysis is performed to differentiate between subjective and objective text. If a group of words carries 'sentiment', then it is subjective; else the phrase is termed objective.
2. Semantic Association: In semantic association, the semantic lexicon related to the subject of the sentence or phrase is defined using grammatical relations and rules of the language. It also included Part Of Speech (POS) tagging, which means to tokenize and categorize the data into various parts of speech using tools such as Natural Language Toolkit (NLTK) [6].
3. Polarity Classification: In polarity classification, subjective phrases/words are classified as positive, negative or neutral based on their determined scores. For example, according to the SentiWordNet, the word 'love' has a positive has a positive score of 0.625; so we can conclude that the phrase 'I love you' contains positive sentiment.

Review of existing literature:-

Sentiment analysis has been a hot topic of research over the past decade and a half. The number of papers published in leading journals about Opinion Mining and Sentiment Analysis has increased exponentially since 2002 [7]. Lay-Ki Soon et al. describe in [8] that in a bulk of scenarios, data pre-processing is vital. Especially when applying sentiment analysis techniques to text derived from social media and micro blogging websites; it is important that the data must first be cleaned and structured. For instance, Apoorv Agarwal et al. in [9] state pre-processing of tweets includes replacing emoticons with sentimental scores using emoticon dictionary, replacing acronyms(say, gr8 with great) and checking for commonplace spelling errors.

Sentiment Analysis and Opinion Mining have been studied at three level of granularity. These are as stated below:

1. Document level: At document level analysis, whole piece of text is classified as positive or negative. And thus, as pointed out by Solanki and Bhumika in [10], it can't be applied to text where sentiment about more than one topics have been expressed.
2. Sentence level: At this level, each sentence is taken into consideration independently, and tagged as positive, negative or neutral. Many researchers, including Hu and Liu in [11], have studies sentiment analysis at the sentence level.
3. Phrase level: Recently, some research has been done to deal with the task of identifying sentiment of text at the phrase level, namely Agarwal et al. (2009) and Wilson et al. (2005), have accomplished substantial research studies at this level of analysis. It is also termed as fine-grained analysis.

However, the endeavors have chiefly been confined to analysis and classification of text in the English language. Relatively minimal effort has been put into sentiment analysis of text in other languages, despite the fact that only 28.6 percent of internet users speak English [12].

However, Vandana et al. in [13] have presented a novel idea of opinion mining of movie reviews written in the Hindi language.

The sources of data used by the researchers and their techniques of obtaining the data vary hugely. The various sources of data collection include social media websites, such as twitter(Neetu and Rajasree, 2013), blogs, movie reviews(Pang and Lee, 2004) etc. The two common methods of obtaining information are web crawlers and APIs. A brief comparison of these two methods is provided in Table 1.

Table 1:- Comparison of APIs and Web Crawlers.

Parameter of evaluation	**API methods**	**Web Crawler**
Implementation	Simple	Complicated
Data retrieved	Structured	Unstructured
Applications	Social media, Micro blogs	All Web resources

A concise review of prominent research done in the field of sentiment analysis is provided in Table 2.

Future Challenges:-

Comprehensive research has been done so far in the field of Sentiment Analysis. However, still certain scope of improvement exists and certain topics along which future research might be directed is enlisted below:

1. Domain Dependence: The current methods of feature based sentiment analysis depend heavily on the domain. As a result, different lexicon databases are being used for different domains. One of the challenges in the future would be to stride towards domain independent classifiers.
2. Word Sense Disambiguation: Mostly sentiment analysis systems determine the polarity without word sense disambiguation. Though some innovative work has been done by Umar Farooq et al. [14] regarding disambiguation methods. In the near future, attempts must be made to weigh in the context of a word.
3. Integrated text and Multimedia Analysis: So far, the research has been primarily carried out in text-based sentiment analysis. But with the explosion of multimedia over the world wide web in recent times, techniques must be developed for mining of sentiments and opinions from multimedia such as images.

Table 2:- Summary of the Survey.

S. No	**Author(s)**	**Title of the Study**	**Data Source(s)**	**Technique Used**	**Performance (Accuracy)**	**Annotations**
1	Mostafa et al	Sentiment Analysis of Social Issues(2012)	Comments from Yahoo! , CNN	Lexicon based approach for document analysis	65%	Explores Sentiment Analysis in social domain
2	Vandana Jha et al	Hindi Opinion Mining System (2015)	Hindi movie reviews from bbc.co.uk/hindi	Unsupervised learning with POS tagging	87.1%	Explores Sentiment Analysis in the language of Hindi

3	Neetu M S and Rajasree	Sentiment Analysis in Twitter(2013)	Tweets extracted using Twitter API	Machine Learning(SVM, Nave Bayes,ME)	NB- 89.5% ME- 90% SVM- 90%	Comparison of various classifiers
4	Wei Yen Chong et al	NLP for Sentiment Analysis (2014)	Tweets collected from twitter.com	Alchemy API, SVM, Decision Tree(J48)	59.85%	Shows how preprocessing is required for better results
5	Yu Huangfu et al	Improved Sentiment Analysis (2015)	Chinese news websites	Logistic Regression Model, Lexicon based approach	Positive-91% Neutral-71% Negative-63.33%	Assigned different sentiment value to title of the text
6	P. Waila et al	Sentiment Analysis of Movie reviews	Self created database and reviews from www.imdb.com	Lexicon based approach using SentiWordNet	77.6%	Uses a novel 'Adverb+Adjective+ Verb' combination
7	Addlight Mukwazvure	Hybrid Approach to Sentiment Analysis	News comments from theguardian.com	SVM classifier, kNN Approach	SVM-73.3% kNN- 74.2%	Shows categorizing text as per domain results in better performance of classifiers
8	Tej Prasad Dhamala el al	A Word Sense Disambiguation Method (2015)	Product Reviews from various websites such as ebay.com	Lexicon based approach for Feature level analysis	87.3%	Introduces concepts regarding Contextual Polarity
9	Owen Rambow et al	Sentiment Analysis of Twitter Data	11,875 tweets from twitter	SVM classifier, WordNet	75%	Used a tree kernel to obviate feature engineering
10	Hu and Bing Liu	Opinion extraction and Summarization (2006)	Product Reviews from amazon.com and epinions.com	Rule based algorithm with WordNet(Lexicon)	84%	Uses different approaches from reviews in different formats
11	Bo Pang and Lee	A Sentimental Education (2004)	Movie Reviews from www.rottentomatoes.com	SVM, Nave Bayes	86.4%	Defines relation between subjectivity detection and polarity classification
12	P.D.Turney	Thumbs Up or Thumbs Down? (2002)	Reviews from Epinions.com	PMI-IR Method	74%	Used algorithm goes beyond analysis of isolated adjectives
13	A. Gamon	Sentiment classification on customer feedback data (2004)	Customer Feedback	Support Vector Machine(SVM)	77.5%	Used a 4-point scale to analyze at document level
14	R.Jose and V.Chooralil	Prediction of Election Result by Enhanced Sentiment Analysis (2015)	Tweets for 3 weeks during Delhi elections	Lexicon based approach using WordNet and SentiWordNet	78.6%	Used word sense disambiguation to achieve improvement of 2.6%

Conclusions:-

Sentiment Analysis is a relatively recent research topic, which was introduced formally in 2001 [2]. Applications of this stream of knowledge discovery are vast and varied. In this paper, we have looked to provide a brief survey of the research done in this field and a holistic review of the basic analysis and classification techniques that are used in

Sentiment Analysis. Furthermore, certain challenges and opportunities for future research work have also been outlined.

References:-

1. B.Pang and L.Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, vol. 2, pp 1-135, Now Publishers, 2008.
2. S.R.Das and M.Y.Chen "Yahoo! for Amazon: Sentiment extraction from small talk on the web", management science, pp 1375-1388, 2001
3. R. Jose and V.S. Chooralil, "Prediction of Election Result by Enhanced Sentiment Analysis on Twitter Data using Word Sense Disambiguation", proceedings of ICCC, pp 638-641 ,2015.
4. Mostafa Karamibekr and A.A. Ghorbani, "Sentiment Analysis of Social Issues", proceedings of International Conference on Social Informatics, pp215-221, 2012.
5. B.Liu, Sentiment analysis and subjectivity, Handbook of Natural Language Processing, 2010.
6. S.Bird, E.Klein and E.Loper, Natural Language Processing with Python, vol.43, 2009.
7. Khairullah Khan, B.B. Baharudin, Aurangzeb Khan and Fazal-e-Malik, "Mining Opinions from Text
8. Documents: A Survey",proceedings of 3rd IEEE International conference on Digital Ecosystems and Technologies, pp 217-222, 2009.
9. Wei Yon Chong, Bhawani Selvaretnam and Lay-Ki Soon, "Natural Language Processing for Sentiment Analysis", proceedings of 4th International Conference on Artificial Intelligence with Application in Engineering and Technology, pp 212-217, 2014.
10. Apoorv Agarwal, Boyi Xie Ilia Vovsha, Owen Rambow and Rebecca Passonneau, "Sentiment Analysis of Twitter Data", Columbia University.
11. Solanki Yogesh Ganeshbhai and Bhumika K. Shah, 978-1-4799-8047-5 IEEE, 2015
12. Minqing Hu and Bing Liu, "Mining Opinion Features in Customer Reviews ", American Association for Artificial Intelligence, 2004.
13. http://www.internetworlstats.com/stat7.htm
14. Vandana Jha, Manjunath N, P.D.Shenoy, Venugopal K.R., L.M.Patnaik, "HOMS: Hindi Opinion Mining System", proceedings of 2nd International Conference on Recent Trends in Information Systems, pp 366-371, 2015.
15. Umar Farooq, Tej Prasad Dhamala, Antoine Nongillard, Yacine Ouzrout and Mohammed Abdul Qadir, "A Word Sense Disambiguation Method for Feature level Sentiment Analysis", proceedings of 9th International Conference on Software, Knowledge, Information Management and Applications, 2015.