Indian Journal of Science and Technology, Vol 8(36), DOI: 10.17485/ijst/2015/v8i36/88609, December 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

# A Text Mining Approach to Extract Opinions from Unstructured Text

**Ananthi Sheshasaayee*** **and R. Jayanthi**

PG and Research Department of Computer Science, Quaid-E-Millath Government College for Women (Autonomous), Chennai - 600002, Tamil Nadu, India; ananthi.research@gmail.com, jayanthi.research@gmail.com

## Abstract

**Background/Objectives:** To extract and interpret public opinion from informal description of text in social media websites. **Methods:** The informal descriptions containing opinions are Tokenized, Parts of Speech Tagging, Word-sense Disambiguation and Text Transformation/Attribute Generation are employed. Sample data pertaining to performance rating of cricket players was collected from twitter, cricinfo and cricbuzz. The reviews were rated against subjective evaluation criteria, the scale ranging from poor, moderate, good, excellent and the linguistic variables were converted into numerical values using fuzzy since the expressed opinions may be easy to understand. **Findings:** Opinions enhance any decision making process, so the influence of opinions from sample data pertaining to performance rating of cricket players are classified as poor, moderate, good and excellent. **Application/Improvements:** Using the proposed approach a generic prototype can be built that can be used to extract opinions and interpret.

**Keywords:** Information Extraction, Opinion Mining, Unstructured Text

## 1. Introduction

Emotions and opinions are the two important logical substance of human life. This in turn produces large data in the social media sites. Opinions can be used to extract valuable information that influences decision making[1]. This article portrays how raw text containing opinions drawn from social media sites are preprocessed, categorized, summarized, evaluated and the influence is furnished to the decision makers. Text preprocessing involves text cleanup, tokenization, part of speech tagging and attribute generation. Text categorization classifies the document collection of many possible user-dependent and application-dependent categories/classes. These categories are generated with distinct characteristics of the opinion to be evaluated. Summarization involves Sentiment classification which uses polarity assignment to express the opinions with positive assessment, negative assessment or neutral assessment. The rest of the paper is organized as follows, Section 2 deals with some existing work on opinion mining, Section 3 discusses the importance of opinion mining and Section 4 presents the Text mining approach to extract opinions from unstructured text. Experimental results are discussed in Section 5 and finally Section 6 concludes the paper by giving the glance and the future direction of research in this area.

## 2. State of Art

The data produced from blogs, discussion forums, social media and social networking sites, reviews are in the form of unstructured text. In[2] suggested text mining techniques are the way to pre-process and extract information from the unstructured text to make better business decision making.

In[3] presents that identity, sharing, conversations, reputation, groups, relationships and presence are the seven building blocks of social media. In[4] opinioned that corporate makes use of social media that offers abundant occasions for gathering user preferences, opinions, ratings about a product or service and assessments.

*Author for correspondence

In[5] suggested that the sentiments attached with a product, the perception about a brand and the perception about new product introduction can be well interpreted using sentiment analysis. In[6] depicted an alternative way of classifying opinions like exciting, irrelevant and objectionable, rather than positive, negative and neutral.

According to[7] sentiment analysis with the help of fuzzy logic deals with reasoning and gives closer views to the exact sentiment values. According to[8] an opinion mining system can be used for both binary and fine-grained sentiment classifications of user reviews. Feature-based sentiment classification is a multistep process that involves preprocessing to remove noise, extraction of features and corresponding descriptors and tagging their property using fuzzy functions.

## 3. Opinion Mining - Assessment of People

Generally speaking, opinion mining or sentiment analysis aims to determine the attitude of the user with respect to some topic or the overall contextual polarity in social media data analysis. The attitude may be his or her judgment or evaluation, affective state, or the intended emotional communication. Natural language processing, text analysis and computational linguistics are used collectively to mine opinions and extract information pertaining to a subject contained in the input source[9]. The focus of sentiment analysis is to classify the polarity of the input text as positive or negative or neutral, at various levels like document, sentence or feature. Sentiments are classified based on emotions expressed in the sentence like "angry", "sad" or "happy".

### 3.1 Approaches to Opinion Mining

Traditional opinion mining approaches can be divided into four main categories:

- **Keyword spotting** is an approach that classifies text based on the affect words like boring, sad, happy or afraid[10].
- **Lexical affinity** detects affect words and in addition it also assigns the "affinity" to a given emotion[11].
- Machine learning methods like support vector machines and Naive Bayes classifier contributes to **Statistical Methods**[12].
- **Concept level** approaches were based on knowledge representation techniques like ontology and semantic networks[13,14].

To extract the context sensitive meaning of a given opinion, the grammatical relationship between words are taken into consideration[15].

## 4. Text Mining Approach to Extract Opinions

The main sources of data are from newswires, reports, official websites related to a particular topic, microblogs which are unstructured or semi-structured generally. The data found may vary in formats. Extracting data and analyzing the extracted data is a time-consuming process and this leads to poor decision making. Classical text mining techniques are applied for extracting information and converting the information to structured data using the suitable categorization methods[16].

Figure 1 shows the application of text mining methods to extract user opinions from social media data with a step-by-step process which includes collecting/gathering, pre-processing, information extraction, text mining operations, evaluation/interpretation and finally decision making. Text Pre-processing involves Noise Removal, Tokenization, Parts of Speech Tagging, Word-sense Disambiguation and Text Transformation/Attribute Generation. "Bag of Words" and "Vector Space Model" are Text representation methods based on feature words and their occurrences, where every word is represented as individual variable with a numeric weight attached to it. Feature Generation and Feature Selection methods based on features contained in a document are used for Information Extraction[17].

The proposed approach makes use of a domain specific feature knowledge base guide the feature extraction process. The fine-grained selection of keywords can be achieved by defining a property P that finds the keywords from the vocabulary of the text that are mostly adjectives.

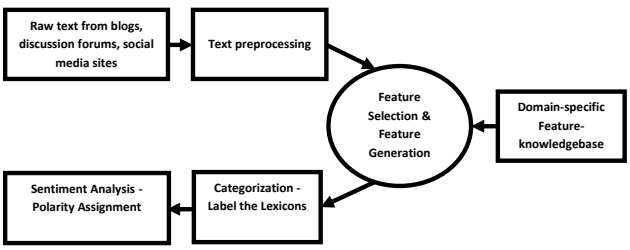


**Figure 1.** Text mining approach to extract opinions from unstructured text.

P(w) is True, if and only if w is from the vocabulary of adjectives V

The mathematical representation of fine-grained selection of keywords $K_s$ is given by:

$$K_s = \{w \mid w \varepsilon V \text{ and } P(w)\}$$

Interpretation/Evaluation - The final step in text mining is Evaluation/Interpretation that brings about either termination or iteration. When the desired results are achieved, the text mining process can be terminated else further iterations can be made to accomplish the required results. The results can be visualized and reports can be presented in a way that is easy for the user to understand using Business Intelligence tools[18,19].

## 5. Experimental Results

Sample data was collected from twitter, cricinfo, cricbuzz. They were pertaining to performance ratings of cricket players. Since data were collected from various sources, they were mostly unstructured. The input data was preprocessed to remove stopwords like exclamatory marks, question marks, emoticons, smilies contained in reviews of the users. Python programming was used to carry out preprocessing. Figure 2 shows the sample screen of the preprocessed data. The first part shows the sentence-level annotations and the second part shows the word-level annotations.

**Figure 2.** Preprocessed data after stop word removal and tokenization.

**Figure 3.** Screenshot of POS tagging using Python and NLTK.

The preprocessed data was subjected to Part-of-Speech tagging with the focus to identify nouns, verbs and adjectives in order to interpret the opinions or comments collected. Python and Natural Language ToolKit(NLTK) were used for this purpose. Figure 3 shows the output of POS tagging.

The extracted reviews were rated against subjective evaluation criteria with support from WORDNET, the scale ranging from poor, moderate, good, excellent and they were represented using the fuzzy set. By changing the linguistic variables into numerical values using fuzzy, the expressed opinions may be represented in a form that is easy to understand[20].

$$F(\text{opinion}) = \{0.2,0.4,0.6,0.8\}$$

Where 0.2 denotes *poor*, 0.4 denotes moderate, 0.6 denotes *good* and 0.8 denotes excellent.

During preprocessing, extra effort had to be applied to establish uniformity of syntax and interpreting the expressed opinion.

## 6. Conclusion

Opinion mining or sentiment analysis is a special field of Text analysis. Text Mining is more improved than data mining since a deeper understanding of the business domain is attained thereby contributes in deriving competitive business intelligence from unstructured data sources. This article aims to extract valuable information

influencing business decisions from raw text. Sentiment analysis is really a challenging area of research for different reviewers. The future work is to fine tune the model and make it generic so that when any opinion is given as input the influence of the opinion on decision making is measured and interpreted.

## 7. References


1. Padmaja S, Fatima SS. Opinion mining and sentiment analysis - An assessment of peoples' belief: A survey. IJASUC. 2013 Feb; 4(1):21–33.
2. Zhou N, Cheng H, Chen H, Xiao S. The framework of text-driven business intelligence. IEEE Wireless Communications, Networking and Mobile Computing WiCom; 2007 Sep. p. 5468–71.
3. Kietzmann JH, Hermkens K, et al. Social media? Get serious! Understanding the functional building blocks of social media. Business Horizons. 2011; 54(3):241–51.
4. Thiel K, et al. Creating usable customer intelligence from social media data: Network analytics meets text mining. KNIME; 2012. p. 1–18.
5. Gundecha P, Liu H. Mining social media: A brief introduction. INFORMS; 2012. ISBN: 978-0-9843378-3-5. Available from: http://dx.doi.org/10.1287/educ.1120.0105
6. Sapno Z, Preeti M. Analyzing Sentiments for Generating Opinions (ASGO) - A new approach. Indian Journal of Science and Technology. 2015 Feb; 8(S4):206–11.
7. Haque AMd, Rahman T. Sentiment analysis by using fuzzy logic. International Journal of Computer Science, Engineering and Information Technology. 2014 Feb; 4(1):33–48.
8. Dalal MK, Zaveri MA. Opinion mining from online user reviews using fuzzy linguistic hedges. Applied Computational Intelligence and Soft Computing. Hindawi Publishing Corporation. 2014. p. 735942. Available from: http://dx.doi.org/10.1155/2014/735942
9. Pang B, Lee L. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. 2008; 2(1-2):1–135.
10. Erik C, Bjorn S, Yunqing X, Catherine H. New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems. 2013; 28(2):15-21. DOI: 10.1109/MIS.2013.30.
11. Andrew O, Clore G, Collins A. The Cognitive Structure of Emotions. Cambridge Univ Press ; 1988.
12. Ryan S, Joseph M, Thomas J. Characterization of the affective norms for English words by discrete emotional categories. Behavior Research Methods. 2007; 39(4):1020–4.
13. Dey L, Haque SKM. Proceedings of the 2nd Workshop on Analytics for Noisy Unstructured Text Data; 2008. p. 83–90.
14. Erik C, Amir H. Sentic computing: Techniques, tools, and applications. Springer; 2012.
15. Kim SM, Hovy EH. Identifying and analyzing judgment opinions. Proceedings of the Human Language Technology/ North American Association of Computational Linguistics; 2006 Jun. p. 200–7.
16. Jayanthi R, Ananthi S. Text mining approach to extract knowledge from social media data to enhance business intelligence. International Journal of Advanced Research in Science and Engineering. 2015 Feb; 4(2):700–2.
17. Ananthi S, Jayanthi R. Exploring the potential of social media data using text mining to augment business intelligence. An International Journal of Advanced Computer Technology. 2014 Apr; 3(4):738–42.
18. Gupta P, et al. Role of text mining in business intelligence. Gian-Jyoti E-journal. 2012; 1(2):1–13.
19. Funk A, et al. Opinion analysis for business intelligence applications. ACM OBI Proceedings of the 1st International Workshop on Ontology-Supported Business Intelligence; 2008. p. 1–7. ISBN: 978-1-60558-219-1. DOI: 10.1145/1452567.1452570.
20. Jusoh S, Alfawareh HM. Applying information extraction and fuzzy sets for opinion mining. ICICIE 2014; Kuala Lumpur. 2014 Jan 15-16. p. 91–4.