Sentiment Analysis: An approach in Natural Language Processing for Data Extraction

Shabina Dhuria

Department of Computer Science, DAV College, Sector-10, Chandigarh

Abstract: Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, web mining, and text mining. Sentiment analysis has been used in several applications including analysis of the repercussions of events in social networks, analysis of opinions about products and services. The growing importance of sentiment analysis coincides with the growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, and social networks. Methods like supervised machine learning and lexical-based approaches are available for measuring sentiments that have a huge volume of opinionated data recorded in digital form for analysis.

Keywords: Sentiment Analysis (SA), Natural Language Processing (NLP), Opinion Mining, Training Set, Classifier.

I. INTRODUCTION

Sentiment Analysis (SA), known as mood extraction [1], is a blooming interest area as an application of Natural Language processing (NLP). Mood Extraction automates the decision making performed by human. It also classifies the polarity of text in terms of positive, negative or neutral (surprise). Based on polarity, a training set is prepared and further classifier is implemented to classify the reviews as positive or negative. Social network revolution plays a decisive role in gathering information containing public opinion. To obtain subjective and factual response from the gathered information, public opinions are extracted by features extractor [20]. Figure 1 shows the layout of predicting the hidden information in relation to user's intensions, likeliness and taste.

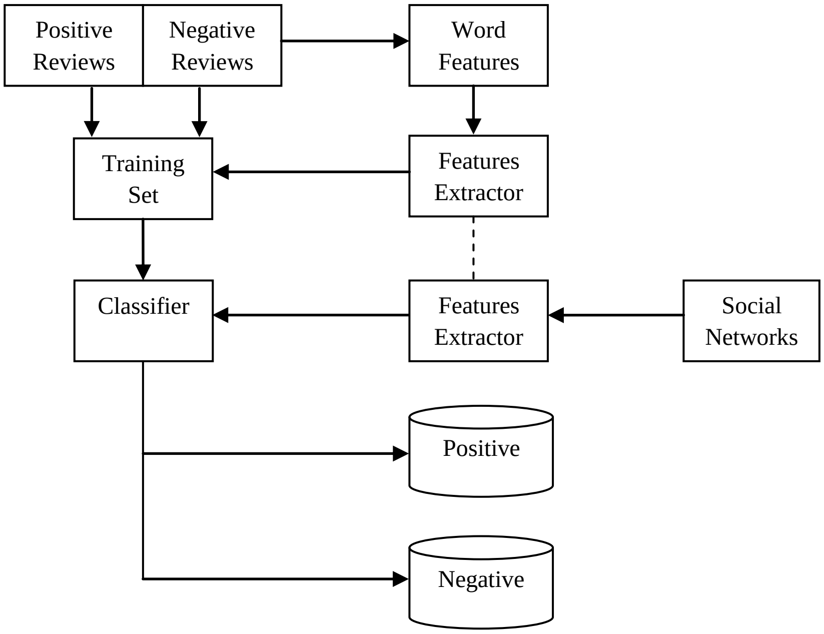

Figure 1: Layout for Sentiment Analysis of data for Social Networks

For this purpose sentiment analysis follows various approaches as discussed below:

- *Subjective lexicon*: It finds nature of word on the basis of a score assigned to each word in a list of words. Nature of word can be positive, negative or objective. The basic techniques used for generating a subjective lexicon are machine translation, using wordnet or Bi-lingual dictionaries [2] [15].
- *Using N-Gram modeling:* An N-Gram model (uni-gram, bi-gram, tri-gram or combination of these) is designed for categorization on the basis of particular training data. This model is trained by using the training data and then, its efficiency is tested using the tested data [4] [9].
- *Machine learning***:** Supervised or Semi-Supervised learning [13] is performed by extracting the features from the text and learn the model. Various in-built tools like Naive Bayes, Support Vector Machine [17] can be used to implement the concept of machine learning [3][10].

Sentiment analysis is also known as Subjectivity analysis. It is the computational study of affect, opinions, and sentiments expressed in text viz. blogs, editorials, newspaper articles and reviews of products, movies, books, etc.

II. Sentiment Analysis

Sentiment analysis classifies the polarity of a given text of the document, sentence or aspect level expressing the opinion as positive, negative or neutral [5]. The sentiment analysis can be performed at one of the following levels:

- *Document-Level Sentiment Classification:* In document level sentiment analysis main challenge is to extract informative text for inferring sentiment of the whole document [6]. Two main approaches for document-level sentiment analysis include supervised learning and unsupervised learning. The supervised approach assumes that there is a finite set of classes into which the document is classified and training data available for each class. The simplest case is composed of two classes viz. positive and negative. Unsupervised approaches are based on determining the Semantic Orientation (SO) of specific phrases within the document for document-level sentiment analysis. If the average SO of these phrases is above some predefined threshold then the document is classified as positive and otherwise it is deemed negative [7].
- *Sentence-Level Sentiment Classification:* The sentiment classification is a fine-grained level than document level sentiment classification in which polarity of the sentence can be given by three categories as positive, negative and neutral. Different types of sentences are handled by different strategies. Sentences that need unique strategies include conditional sentences, question sentences and sarcastic sentences [6]. Sarcasm is extremely difficult to detect and it exists mainly in political contexts [7].
- *Aspect-Based Sentiment Analysis:* The above two approaches work with either the whole document or each individual sentence. In many cases entities have many aspects (attributes) and each of the aspects have a different opinion. This happens in reviews about products or in discussion forums related to specific product categories (such as cars, cameras, smart phones, and even pharmaceutical drugs) [8]. Aspect-based sentiment analysis, also called feature-based sentiment analysis, focuses on the recognition of all sentiment expressions within a given document and the aspects to which they refer [7].

III. Applications areas of Sentiment Analysis

Sentiment Analysis or Opinion Mining is basically used for determining the subjective nature of the data. The domains where Sentiment Analysis is used are as follows:

- *Aid in decision making:* Decision making is an important part of new life. It ranges from “which car to buy”, “which cafe to go” and “which tourist place to visit”. The reviews given by old customers of a particular product are processed by Sentiment Analysis and a best case answer is provided to the user [16].
- *Improving the Quality of the Products:* For every product, there is series of manufacturing firms which leads to a tough competition. Firms use Sentiment Analysis for the better analysis of product. The reviews and opinions of customers are used to improve the quality of product. This concept also leads to the development of innovative products [11].

- *Recommendation Systems:* It is provided to the users for providing their views. This system also provides the development of a great corpus. There are numerous websites with an in-built recommendation system. These types of websites are generally related to the books, music, online media, and film industry. Recommendation system also maintains some important information of user like personal information likes and dislikes previous history and his friend's information to provide more suggestions [12].
- *Business Strategies:* Developing a strategy for business is not the work of an individual, but a team work. This team includes the higher authorities, experts, developers, junior staff and the most important is the customers. Now, the issue arises, how to communicate with the customers for their assistance. Sentiment analysis used the response of the customers, their needs and demands to generate a future strategy and cover the previous flaws.
- *Business Intelligence:* Sentiment analysis is used to search the web for opinions and reviews of these opinions from different Blogs, Amazon, tweets, etc. It also helps in Brand analysis or competitive intelligence, new product perception, product and service benchmark and market forecasting [12].
- *Political SA:* It has numerous applications and possibilities viz. analyzing trends, identifying ideological bias, targeting advertising or messages, gauging reactions, etc. It is also useful in evaluation of public opinions and views or discussions of policy.
- *SA and Sociology:* Idea propagation through groups is an important concept in sociology. Opinions and reactions to ideas are relevant to adoption of new ideas and analyzing sentiment reactions on blogs can give insight to this process e.g. modeling trust and influence in the blogosphere using link polarity [18].
- *SA and Psychology:* It has potential to augment psychological investigations or experiments with data extracted from natural language text.

IV. CHALLENGES TO SENTIMENT ANALYSIS

Sentiment Analysis is the computational study of affect, opinions, and sentiments expressed in text viz. blogs, editorials, newspaper articles and reviews of products, movies and books. General challenges in the research of sentiment analysis are:

- *Noise (abbreviations, slangs):* Noise on the web is increasing day by day. Abbreviations, slangs, emotions are commonly used by people for ease of use but for language processing, these contribute towards the increase in complexity. For example: tour ws awsmmm. This type of errors leads to spelling variations. For example: awesome word can be found as awesm, awsumm, awsuum.
- *Unstructured Data***:** Web contains a large amount of unstructured data. Same entity is represented by different forms [21]. The sources of web varies from web documents, journals, books, health records, internal files of an industry, companies logs, multimedia platforms, texts, videos, audios, images etc. So, this diversity in the sources of data and different formats increases the complexity [18].
- *Contextual Information:* Actual sense of the text varies from domain to domain; this property is referred as contextual property. So, based on the context, the behaviour of the word changes.
- *Word Sense Disambiguation:* One word may have multiple meanings. This concept also affects the polarity of the word. For example-In English word "good" have multiple senses according to the usage in a particular sentence [15].
- *Language Constructs:* Different styles in a language lead to different challenges. Some of the challenges while dealing with English language are as under:

Table 1: English Sentence - word order

Sham ate two mangoes	Correct order SVO
Ate sham two mangoes	Incorrect order VSO
Two mangoes ate ram	Incorrect order OVS

- *Word order:* for identifying the subjective nature of the text, arrangements of words in a sentence play an important role. In English language, there is a fixed order set by grammatical rules i.e. subject is followed by verb which is further followed by object. This concept is explained by examples in Table.

- *Morphological Variations:* The concept of morphological variables states that information is fused in the words. Table 1 shows the verb 'ate' which carries much information apart from just the root word. For example: Smith ate apple.
  Smith is eating apple.
  Smith eats apple.

Thus with the variations in the root word, there can be many words.

- *Handling Spelling Variations:* As in Punjabi language, one word can possess many spellings, so this lead to high complexity. It becomes complex to process all the variants a single word. This problem is also faced during training the model.
- *Lack of resources:* Lack of tools, resources, corpora lead to great struggle while doing sentiment analysis for Indian languages.

V. Conclusion

Sentiment Analysis can be used for analyzing opinions in blogs, articles, Product reviews, Social Media websites, Movie-Review websites where a third person can narrates the views. It has many applications and is an important field to study. It has strong commercial interest because companies want to know how their products are being perceived and also prospective consumers can explore the thinking of existing users. Different types of features and classification algorithms may be combined for efficient analysis. Applications of sentiment analysis and challenges are also discussed in the paper. Providing satisfactory solutions to these challenges will make the area of sentiment analysis more widespread.

References

[1] "Case Study: Advanced Sentiment Analysis". Retrieved 18 October 2013.

[2] Bing Liu, "Sentiment Analysis and Subjectivity". Handbook of Natural Language Processing, Second Edition, 2010.

[3] "Sentiment Analysis on Reddit". Retrieved 10 October 2014.

[4] G.Vinodhini, RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, 2(6), June 2012.

[5] B. Arti, Chandak M. B. and Z. Akshay, "Opinion Mining and Analysis: A Survey", International Journal on Natural Language Computing (IJNLC), 2(3), June 2013.

[6] Jagtap V. S. and Pawar Karishma, "Analysis of different approaches to Sentence-Level Sentiment Classification", International Journal of Scientific Engineering and Technology, 2(3), pp. 164-170, April 2013.

[7] F. Ronen, "Techniques and Applications for Sentiment Analysis", Communications of the ACM, 56(4), pp. 82-89, April 2013.

[8] P. Rudy and Th. Mike, "Sentiment analysis: A combined approach" , International Journal of Informatics, 3, pp. 143–157, 2009.

[9] H.Taneja, S.Dhuria and K.Sukhija "Natural Language Processing: A Backbone for Computational Linguistics", DHE Sponsored National Conference on Computational Sanskrit – Issues and Challenges, pp. 187-190, Dec. 2013.

[10] Shabina Dhuria and Harmunish Taneja, "A Survey on Sentiment Analysis and Opinion Mining", International Conference on Computing Sciences, pp. 124-128, Nov. 2013, Elsevier.

[11] K.R. Chowdhary, "Natural Language Processing", M.B.M. Engineering College, Jodhpur, India April 29, 2012.

[12] Xiaoyong Liu, "Natural Language Processing", School of Information Studies at Syracuse University.

[13] Pang B, Lee L and Vaithyanathan S. Thumbs up? Sentiment Classification using machine learning techniques. In: Proceedings of EMNLP-02, 7th Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA, Association for Computational Linguistics, pp.79–86, 2002.

[14] Pang B and Lee L, "Opinion mining and sentiment analysis: Foundations and Trends in Information Retrieval", pp.1–135, 2008.

[15] Melville and Wojciech Gryc, "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification" 2009.

[16] Zhai Zhongwu, Liu Bing, Xu Hua and Xu Hua, "Clustering Product Features for Opinion Mining." WSDM'11, Feb. 2011.

[17] Chaovalit and Lina Zhou, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches", In: Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.

[18] Mikhail Bautin, Lohit Vijayarenu and Steven Skiena, "Sentiment analysis for news and blogs", In: Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2008.

[19] Shabina Dhuria and Harmunish Taneja, "Ontology Equipped Natural Language Processing for Real World Applications", International Journal of Advanced Research in Computer Science and Software Engineering, 4(4), 2014

[20] Pushpa R. Suri and Harmunish Taneja, "Web Objects Clustering Through Aggregation for Enhanced Search Results", International Journal of Scientific & Engineering Research, 2(8), August 2011

[21] Pushpa R. Suri and Harmunish Taneja, "Object Oriented Information Computing over WWW", International Journal of Computer Science Issues, 7(3), May 2010