A Cognitive Study of Sentiment Analysis Techniques and Tools: A Survey

[1]Upma Kumari, [2]Dinesh Soni, [3]Dr. Arvind K Sharma
[1]Dept. of CSE, Rajasthan Technical University Kota, Rajasthan, India
[2]Rajasthan Technical University Kota, Rajasthan, India
[3]Dept. of CSI, University of Kota, Kota, Rajasthan, India

Abstract

Sentiment analysis is a task of identifying positive and negative opinion, emotion and evaluation in text available over the social networking websites and the World Wide Web. The sentiment analysis has been gained quite popularity in the recent years. The analysis serves as an important feedback for further improvement in the offered services and user experiences. Several techniques have been utilised frequently including machine learning approaches and vocabulary oriented semantic algorithms. This paper presents a cognitive study of various techniques and tools which have been used in the sentiment analysis process.

Keywords

Sentiment Analysis, Machine Learning, Tools, Techniques

I. Introduction

The growth of the web and social networking sites such as Facebook, Instagram, Twitter, Blogs, and Forums etc. have been emerged into a huge volume of user reviews and opinions about particular aspects of products or services. People like to share their experiences, thoughts, opinions, feelings, and preferences according to their understanding and observation about the services. Their point of view or impression may be positive, negative or neutral. This opinion is used for identifying trends, user interest, and prediction of stock markets, political polls, and market researches, enhancing the user experience by presenting the things of their own interest and to influence them towards a particular direction. For one particular aspect, one may have a positive opinion while some other may have a negative opinion at the same time. Thus, classifying opinion and sentiment of people is a difficult task. Furthermore, the shared reviews and feelings are not in specifically structured format, thus identifying its positivity or negativity perspective automatically, is also convenient. Therefore, analysis of an unstructured format of text and extract the information for determining the user's sentiments requires special machine learning techniques and semantic algorithms for their classification. In sentiment analysis, major tasks listed are subjectivity and sentiment classification, sentiment lexicon generation, opinion spam detection and quality of reviews [2].

The rest of the paper is organised as follows: Section II explains the complete process of sentiment analysis. Section III presents literature survey. Section IV shows sentiment analysis techniques and their taxonomy. Section V explores different tools. Section VI provides proposed methodology. Section VIII concludes the paper while references are mentioned in the last.

II. Sentiment Analysis Process

Sentiment analysis is a process of finding user's review towards a website or a product. Sentiment analysis is classified into positive comment, negative comment or neutral comment. Fig. 1 shows the complete process of sentiment analysis that refers how the input is being classified on the various steps.

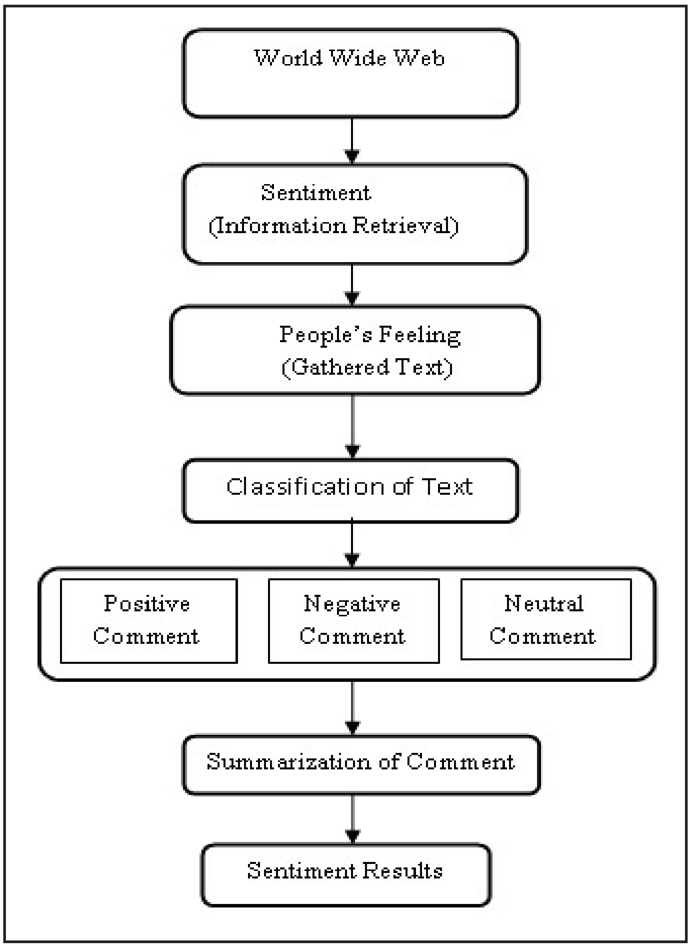

Fig. 1: Process of Sentiment Analysis

The sentiment analysis summarization process contains three main steps, first is Sentiment information retrieval, second is Sentiment classification and third is Sentiment summarization. Review text is retrieved from review websites such as Twitter, Facebook, Amazon and News Sites etc. Sentiment text in the blog, reviews, comments, microblogs etc. contains subjective information about the topic or issue. Sentiment results are generated based on features (sentiment sentences) selection about a matter.

II. Literature Survey

In this section, a literature survey on the basis of literature and research papers of past year's carried out by several researchers in the sentiment analysis domain has been presented. A few work are has been discussed.

In 2016, Fang Luo et al. [1], proposed a method that could perform at the sentence level and document level while it failed at word level sentiment analysis.

In 2016, Ebru Aydogan et al. [2], has carried out a comprehensive survey on Sentiment Analysis using different types of Machine learning algorithms and found that SVM and NB were most commonly employed due to their higher estimation capability.

In 2016, Mohammed Qasem et al. [3], used two weighting schemes namely Unigram term frequency (TF) and Bigram term frequency-inverse term frequency (TF-IDF) to classify different tweets into positive, negative and neutral classes. Positive tweets were determined by positive emoticons and negative tweets by negative emoticons while the neutral tweets were defined as those with no emoticons or keywords that indicate polarity, like happy, sad, good, bad, etc. The weak point was due to the automatic annotation of neutral class. In 2016, M. Trupthi et al. [4], explored machine learning approaches with different feature selection schemes, to identify the best possible approach and found that the classification using high information features, resulted in more accuracy than Bigram Collocation. They also proposed that there was a scope for improvement using hybrid techniques with various classification algorithms. In 2016, Orestes Appel et al. [5], proposed a hybrid system using Naive Bayes (NB) and Maximum Entropy (ME) methods to the same dataset which worked very well with the high level of accuracy and precision.

In 2016, S. Brindha et al. [6], presented a survey on different classification techniques (NB, KNN, SVM, DT, and Regression). Authors found that almost all classification techniques were suited to the characteristics of text data. Authors concluded that further study on classification development could get the enhanced quality of text results and accurate data along with minimized accessing time. In 2015, Huang Zou et al. [7], introduced a syntactic feature in pre-existing words-bag method that revealed more on Pos Tags. They applied SVM and Naive Bayes and observed that word dependency and pos tags did improved accuracy. In 2014, P. Kalaivani and K.L. Shunmuganathan [8], proposed an improved KNN algorithm by incorporating information gain for feature selection to improve and show that this approach out performed Naive Bayes and KNN.

In 2013, Mostafa Karamibekr and Ali A. Ghorbaniss [9], focused their work mainly on topic originated opinion mining, where only opinions about particular topic or issue would be considered. A text may not necessarily contain opinion about a targeted topic. As per their experiments the Precision or Recall method for subjectivity of classification at the sentence level was considerably lower than those of previous works. However, the F-measure was higher which indicated an improved overall balance between Precision and Recall. In 2012, Saif et al. [10], worked on Semantic sentiment analysis of twitter and their results shown that the semantic feature model out performs the Unigram and POS baseline for identifying both negative and positive sentiment. In 2010, Alexander Pak and Patrick Paroubek [11], worked on Corpus for sentiment analysis using twitter that is most popular Microblogging platform. Authors presented a method for an automatic collection of a corpus that could be used to train a sentiment classifier. In 2010, Khin Phyu Phyu Shein et al. [12], proposed combination of using Natural Language Processing techniques (NLP), ontology based on Formal Concept Analysis (FCA) design, and Support Vector Machine (SVM) which have used for classifying the software reviews as positive, negative or neutral. In 2009, Sun Yueheng et al. [13], proposed a method through which they could decide the 'mood' of a user review. And to do this they took an approach for automatic sentiment analysis by:

- Generating positive and negative sentiment words from Tongyici Cilin with paradigm words.
- Determining the sentiment orientation of ambiguous words according to their contexts.
- Setting up proper weight factors to different part-of-speeches.
- Expanding the initial sentiment words by an iterative process, but through this, they could only achieve an average precision of 83.52%.

In 2009, Cheng Mingzhi at el. [14], proposed a method in which a word association graph was constructed from a text corpus, i.e. product reviews, in which each node was a word and if there was an edge between two words, it meant two words co-occur in the same sentence. And then, with a random walk algorithm, the sentiment score was calculated for all the words in the graph at one time. In 2007, Mostafa Al Masum Shaikh et al. [15], presented a paper in which various approaches to sense sentiments contained in a sentence by applying a numerical-valence based analysis.

IV. Sentiment Analysis Techniques

Sentiment analysis is the process of classifying the opinions conveyed in the documents or statements of the web contents as positive, negative or neutral. The huge amount of data available on the Internet which is right now useless, so to make that data useful we need to convert that data through sentiment analysis process. Sentiment analysis concept or Opinion Mining refers the use of natural language processing, text analysis and computational linguistics to identify and get subjective information in source materials. Sentiment analysis is frequently available for reviews and social media for different applications. Generally, sentiment analysis means to determine the idea of a speaker or a writer or user regarding a topic or the overall explanation of a document. The approach may be person's evaluation, affective state (the sentiments of the author during writing) or the intended emotional communication (the emotional effect the author wishes for a reader). Several techniques have been presented in the recent years, some rely on the machine learning approaches with supervised, unsupervised or semi-supervised learning and other may use semantic-based approaches. Moreover, few hybrid approaches may also be used from techniques related to different domains.

The taxonomy of sentiment analysis techniques is classified into the following categories as shown in fig. 2.

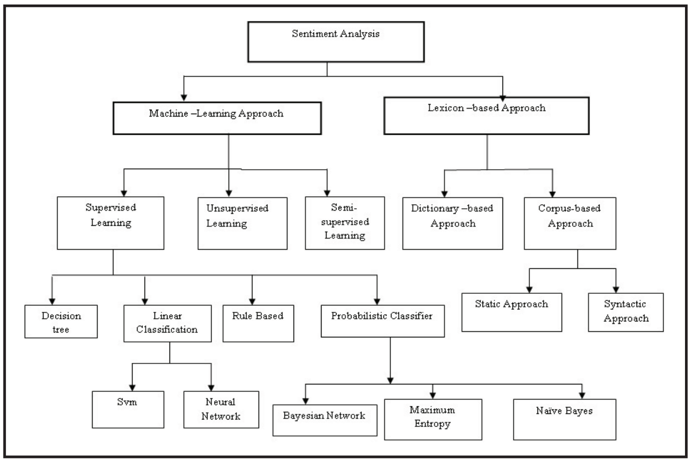

Fig. 2: Taxonomy of Sentiment Analysis Techniques

A. Supervised Learning

It is the machine type learning task of generalizing a function from tagged training data. Supervised learning method is a successful solution in classification and has been used for sentiment classification with very promising results. Supervised classification techniques such as Naive Bayes, SVM, DT, KNN and Regression are used to find the classification accuracy for different dataset [6].

B. Unsupervised Learning

Text classification classifies documents into a number of predefined categories. A large number of training data are used for categorization in supervised learning and its performance depends on the similarity between the training and the testing data. In text categorization, it is usually difficult to create labelled training documents and it requires human efforts. However, it is easy to collect the unlabeled documents. The unsupervised learning methods can overcome these difficulties. Many research studies are presented in this field because in sentiment classification unsupervised learning is less dependent on the domain or topic and its performance is much better. Tweet words are used as features and tweet data were accumulated in positive, negative and neutral [2].

C. Semi-supervised Learning:

Semi-supervised learning is a little bit different from supervised learning and unsupervised learning. Semi-supervised learning uses both labelled and unlabeled data. Main idea behind the semi supervised learning is that unlabeled data hold many information about classes, but they contain information about joint distribution over classification features. To improve the results of supervised and unsupervised learning, where there is limited labelled data we use semi-supervised learning method [2].

V. Sentiment Analysis Tools

Now day's several sentiment analysis tools are available in the market. The (open-source text analytics tools) have been used for natural language processing (NLP), such as information extraction and classification which can also be applied for web sentiment analysis, In this section we have been explored some of the popular tools used for sentiment analysis. Some of them as are follows:

A. Ntlk

Natural Language Analyses with NLTK

A natural language toolkit is a tool for text processing, cataloging, tokenization, stopping, tagging, parsing, etc. It provides easy-to-use interfaces to more than 50 corpora and lexical resources such as WordNet [17].

B. Opinion Finder

It supports in identifying individual sentences and to create different parts of subjectivity in these sentences, it includes the conclusion holder of the subjectivity and words that are incorporated into expressions by communicating positive or negative suppositions [18].

C. Open Nlp

openNLP™

The toolkit which is based on machine learning technique, and is used for processing natural language text is the Apache OpenNLP library. The most common NLP task included are tokenizer, part of speech tagger, named entity extractor, chunker, parser, and conference resolution. In order to build more advanced text processing services, these tasks are usually required [19].

D. Web Fountain

WF WebFountain WebDesign

It is a sentiment analysis tool that completes the requirements of analysis agents such as text gathering, storing indexing and querying. At distributed platforms, this high-performance tool can be used [20].

E. Weka

WEKA The University of Waikato

This tool is based on machine learning techniques, JAVA programming language is used to implement this tool, and it has its GUI to show the data. Many algorithms and techniques are used such as Classification, clustering, preprocessing, linear regression[21].

F. Ling Pipe

Ling pipe is used for linguistic processing for text, including clustering, cataloging, and extraction [22].

G. Opinion Observer

This tool is used for analyzing and comparing the opinions, which are user generated contents on the internet. As well as it shows the resulting graphical format with respect to opinion generated for product feature by feature. It uses WordNet exploring method to assign prior polarity.

H. Review Seer Tool

The work done by the Review site is automated by this tool. To collect positive and negative sentiment for assigning a score to extracted feature terms, we have to use the Naive Bayes classifier approach.

I. Red Opals

It is a tool that enables the users to determine the opinion orientations of products based on their features. It assigns the score to each product based on feature extracted from the customer reviews.

J. Stanford Parser

It is used as a pos tagger and sentence parsing from the NLP group [16].

VI. Proposed Methodology

In this section, the proposed methodology has been presented with following steps:

A. Data Collection:

- Defining dictionary of positive and negative adjectives
- Defining polarity of positive and negative adjectives
- Defining canonical tagging

B. Data Processing

- Extracting of useful data from social networking web sites.
- Removing ambiguous data e.g., sarcasm, interrogative comments.
- Integrating multiple social media channels for accuracy.
- Extracting sentiments from comments.

C. Decision Making

- Use of polarity databases and comments as decision making for sentiment analysis.

D. Data Sources

We discuss the various important data sources to be used for sentiment analysis. There are many data sources available on the World Wide Web e.g. Blogs, Microblogs, Forum Sites, Review Sites, Datasets , Posts, and Corpus etc. from there, the data can be determined in the form of speech, text etc.

1. Blogs

Blogs are the reviews about a particular topic, event or issues in which people express their emotions in their own ways over the internet [16].

2. Micro Blogs

Microblogging platforms are used by different people to express their opinion about different topics, thus it is a valuable source of people's opinion [11] twitter.

3. Forums Sites

An internet forum, or message board, is an online discussion website where people can hold a conversation in the form of posted message.

4. Review Sites

Review sites are the websites which consider many reviews of the customers in order to provide best products and services. Many review sites are available such as www.myntra.com,www.yahoo.com etc.

5. Datasets and Posts

We can take dataset from many websites like News sites, Social media websites and many other websites for sentiment analysis. In the online posts, we share videos and photos and sometimes likes and dislikes some posts.

6. Corpus

Using Twitter API we collect a corpus of text, and form a dataset of three classes, positive sentiments, negative sentiments, and a set of objective texts [1].

VII. Conclusion

Now a day's Sentiment analysis is one of the hot research areas for the researchers. The information gathered from the online data sources like blogs, microblog, forums, review sites etc. has been playing an important role in expressing people's feelings, thoughts, emotions, and opinions for the particular topic, event or issue.

The objective of sentiment analysis is mining the opinion behind the user's statement and revealing the user's interest, preferences and thoughts about the particular thing.

In this paper a cognitive study of sentiment analysis techniques and tools has been presented. The proposed methodology provides important phases the sentiment of text, whether it is positive or negative. This paper will be helpful to the researchers of the sentiment analysis domain.

Reference

[1] Fang Luo et al.,"Affective-feature-based Sentiment Analysis using SVM Classifier", IEEE 20th International Conference on Computer Supported Cooperative Work in design 2016.

[2] Ebru Aydoan et al.,"A Comprehensive Survey for Sentiment Analysis Tasks Using Machine Learning Techniques", IEEE, 2016.

[3] Mohammed Qasem et al.,"Twitter Sentiment Classification Using Machine Learning Techniques for Stock Markets", IEEE, 2015.

[4] M. Trupthi et al.,"Improved Feature Extraction and Classification - Sentiment Analysis,"International Conference on Advances in Human Machine Interaction (HMI-2016), March 03-05, 2016, R. L. Jalappa Institute of Technology, Doddaballapur, Bangalore, India.

[5] Orestes Apple et al.,"A Hybrid Approach to Sentiment Analysis", IEEE, 2016.

[6] S. Brindhaet et al.,"A Survey on Classification Techniques for Text Mining", 3rd International Conference on Advanced Computing and Communication Systems (ICACCS-2016), Jan. 22-23, 2016, Coimbatore, INDIA.

[7] Huang Zou et al.,"Sentiment Classification Using Machine Learning Techniques with Syntax Features", International Conference on Computational Science and Computational intelligence, IEEE, 2015.

[8] P. Kalaivani et al.,"An Improved K-nearest-neighbor algorithm using Genetic Algorithm for Sentiment Classification" International Conference on Circuit, Power and Computing Technologies [ICCPCT], IEEE, 2014.

[9] Mostafa Karamibekr et al.,"A Structure for Opinion in Social Domains", IEEE, 2013.

[10] Hassan Saif et al.,"Semantic Sentiment Analysis of Twitter", 11th International Semantic Web Conference (ISWC 2012).

[11] Alexander Pak et al.,"Twitter as a Corpus for Sentiment Analysis and Opinion Mining", 2010.

[12] Khin Phyu Phyu Shein et al.,"Sentiment Classification based on Ontology and SVM Classifier", 2010 Second International Conference on Communication Software and Networks IEEE, 2010.

[13] Sun Yueheng et al.,"Automatic Sentiment Analysis for Web User Reviews", 1st International Conference on Information Science and Engineering (ICISE2009).

[14] Cheng Mingzhi et al.,"A Random Walk Method for Sentiment Classification", Second International Conference on Future Information Technology and Management Engineering, IEEE, 2009.

[15] Mostafa Al Masum Shaikh et al.,"An Analytical Approach to Assess Sentiment of Text", IEEE, 2007.

[16] Dr. Arvind K Sharma et al.,"Web Opinion Mining Techniques and Tools for Finding User's Opinion", In the 5th International Conference on Innovative Research in Engineering Science and Management (ICIRESM-16) at India International Centre, Max Mueller MargS, New Delhi, India.

[17] [Online] Available: http://Ntlk.logo

[18] [Online] Available: http://opinion finder. Logo/

[19] [Online] Available: http://open nlp.biolab.si/

[20] [Online] Available: http://webfountain.logo/

[21] [Online]http://www.cs.waikato.ac.nz/ml/weka

[22] [Online] http://alias-i.com/lingpipe.

Upma Kumari has received her B.Tech degree in Computer Science and Engineering from Modi Institute of Technology, affiliated to Rajasthan Technical University, Kota, India in 2015 and presently pursuing her M.Tech Computer Science and Engineering from Rajasthan Technical University, Kota, Rajasthan, India. Her research interest includes-Sentiment Analysis, Opinion Mining, and Machine Learning.

Dinesh Soni has received his B.E. degree in Computer Engineering from Rajasthan University, Jaipur, India, in 2007 and M.Tech degree in Computer Technology from Indian Institute of Technology, Delhi, India in 2015. He is working as Assistant Professor in Department of Computer Science and Engineering, Rajasthan Technical University, Kota since 2008. His research interests include: machine learning and Computer Vision.

Dr. Arvind K Sharma has received his Ph.D degree in Computer Science in the year 2013. He has more than 13 years of work experience in academic field. He has published more than 40 Papers in many National, International Journals and Conference Proceedings. He is a Managing Editor of International Journal of Computer Science and Technology. He is also Editorial Board Member and Reviewer of several National and International Journals and Conferences. He is a Member of numerous academic and professional bodies such as IEEE, WASET, IEDRC, IAENG Hong Kong, IACSIT Singapore, UACEE UK, ACM, New York. His area of interest includes- Web Usage Mining, Web Engineering, Opinion Mining, Data Analytics and Machine Learning Tools.