International Journal of Research in Engineering, Science and Management
Volume-2, Issue-8, August-2019
www.ijresm.com | ISSN (Online): 2581-5792

A Survey on Challenges and Analysis of Sentimental Approaches using Data Mining

S. Safeena[1], R. Manicka Chezian[2]

[1]*Research Scholar, Dept. of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India*

[2]*Associate Professor, Dept. of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India*

***Abstract*: Sentiment Analysis (SA) is an ongoing field of research in text mining field. SA is the computational treatment of opinions, sentiments and subjectivity of text. This survey paper tackles a comprehensive overview of the last update in this field. Many recently proposed algorithms' enhancements and various SA applications are investigated and presented briefly in this survey. These articles are categorized according to their contributions in the various SA techniques. The related fields to SA (transfer learning, emotion detection, and building resources) that attracted researchers recently are discussed. The main target of this survey is to give nearly full image of SA techniques and the related fields with brief details. The main contributions of this paper include the sophisticated categorizations of a large number of recent articles and the illustration of the recent trend of research in the sentiment analysis and its related areas.**

***Keywords*: data mining, sentiment analysis**

1. Introduction

Sentiment Analysis (SA) or Opinion Mining (OM) is the computational study of people's opinions, attitudes and emotions toward an entity. The entity can represent individuals, events or topics. These topics are most likely to be covered by reviews. The two expressions SA or OM are interchangeable.

They express a mutual meaning. However, some researchers stated that OM and SA have slightly different notions. Opinion Mining extracts and analyzes people's opinion about an entity while Sentiment Analysis identifies the sentiment expressed in a text then analyzes it. Therefore, the target of SA is to find opinions, identify the sentiments they express, and then classify their polarity as shown in Fig. 1.

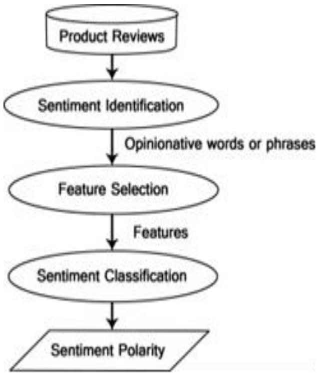

Fig. 1. Sentiment analysis process on product reviews

Sentiment Analysis can be considered a classification process as illustrated in Fig. 1. There are three main classification levels in SA: document-level, sentence-level, and aspect-level SA. Document-level SA aims to classify an opinion document as expressing a positive or negative opinion or sentiment. It considers the whole document a basic information unit (talking about one topic). Sentence-level SA aims to classify sentiment expressed in each sentence. The first step is to identify whether the sentence is subjective or objective. If the sentence is subjective, Sentence-level SA will determine whether the sentence expresses positive or negative opinions. Wilson et al. have pointed out that sentiment expressions are not necessarily subjective in nature. However, there is no fundamental difference between document and sentence level classifications because sentences are just short documents There are many applications and enhancements on SA algorithms that were proposed in the last few years. This survey aims to give a closer look on these enhancements and to summarize and categorize some articles presented in this field according to the various SA techniques. The authors have collected fifty-four articles which presented important enhancements to the SA field lately. These articles cover a wide variety of SA fields. They were all published in the last few years. They are categorized according to the target of the article illustrating the algorithms and data used in their work. According to Fig. 1, the authors have discussed the Feature Selection (FS) techniques in details along with their related articles referring to some originating references. The Sentiment Classification (SC) techniques, as shown in Fig. 2, are discussed with more details illustrating related articles and originating references as well.

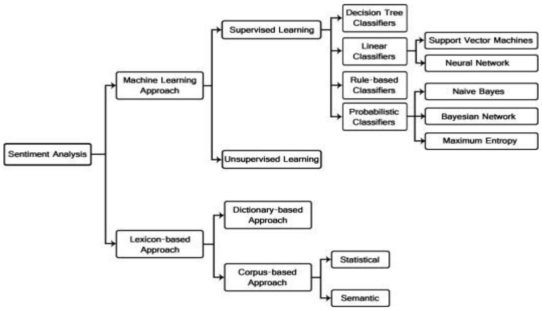

Fig. 2. Sentiment classification techniques

This survey can be useful for new comer researchers in this field as it covers the most famous SA techniques and applications in one research paper. This survey uniquely gives a refined categorization to the various SA techniques which is not found in other surveys. It discusses also new related fields in SA which have attracted the researchers lately and their corresponding articles. These fields include Emotion Detection (ED), Building Resources (BR) and Transfer Learning (TL). Emotion detection aims to extract and analyze emotions, while the emotions could be explicit or implicit in the sentences. Transfer learning or Cross-Domain classification is concerned with analyzing data from one domain and then using the results in a target domain. Building Resources aims at creating lexica, corpora in which opinion expressions are annotated according to their polarity, and sometimes dictionaries. In this paper, the authors give a closer look on these fields.

There are numerous number of articles presented every year in the SA fields. The number of articles is increasing through years. This creates a need to have survey papers that summarize the recent research trends and directions of SA. The reader can find some sophisticated and detailed surveys including. Those surveys have discussed the problem of SA from the applications point of view not from the SA techniques point of view.

Two long and detailed surveys were presented by Pang and Lee and Liu They focused on the applications and challenges in SA. They mentioned the techniques used to solve each problem in SA.

2. Feature selection in sentiment classification

Sentiment Analysis task is considered a sentiment classification problem. The first step in the SC problem is to extract and select text features. Some of the current features are

Terms presence and frequency: These features are individual words or word n-grams and their frequency counts. It either gives the words binary weighting (zero if the word appears, or one if otherwise) or uses term frequency weights to indicate the relative importance of features

Parts of speech (POS): finding adjectives, as they are important indicators of opinions.

Opinion words and phrases: these are words commonly used to express opinions including good or bad, like or hate. On the other hand, some phrases express opinions without using opinion words. For example: cost me an arm and a leg.

Negations: the appearance of negative words may change the opinion orientation like not good is equivalent to bad.

A. Feature selection methods

Feature Selection methods can be divided into lexicon-based methods that need human annotation, and statistical methods which are automatic methods that are more frequently used. Lexicon-based approaches usually begin with a small set of 'seed' words. Then they bootstrap this set through synonym detection or on-line resources to obtain a larger lexicon. This proved to have many difficulties as reported by Whitelaw et al. Statistical approaches, on the other hand, are fully automatic.

The feature selection techniques treat the documents either as group of words (Bag of Words (BOWs)), or as a string which retains the sequence of words in the document. BOW is used more often because of its simplicity for the classification process. The most common feature selection step is the removal of stop-words and stemming (returning the word to its stem or root i.e. flies → fly).

In the next subsections, we present three of the most frequently used statistical methods in FS and their related articles. There are other methods used in FS like information gain and Gini index.

1) *Point-wise Mutual Information (PMI)*

The mutual information measure provides a formal way to model the mutual information between the features and the classes. This measure was derived from the information theory. The point-wise mutual information (PMI) $M_i(w)$ between the word w and the class i is defined on the basis of the level of co-occurrence between the class i and word w. The expected co-occurrence of class i and word w, on the basis of mutual independence, is given by $P_i \cdot F(w)$, and the true co-occurrence is given by $F(w) \cdot p_i(w)$.

The mutual information is defined in terms of the ratio between these two values and is given by the following equation:

Mi(w)=logF(w)·pi(w)F(w)·Pi=logpi(w)Pi (1)

The word w is positively correlated to the class i, when $M_i(w)$ is greater than 0. The word w is negatively correlated to the class i when $M_i(w)$ is less than 0.

PMI is used in many applications, and there are some enhancements applied to it. PMI considers only the co-occurrence strength. Yu and Wu [4] have extended the basic PMI by developing a contextual entropy model to expand a set of seed words generated from a small corpus of stock market news articles. Their contextual entropy model measures the similarity between two words by comparing their contextual distributions using an entropy measure, allowing for the discovery of words similar to the seed words. Once the seed words have been expanded, both the seed words and expanded words are used to classify the sentiment of the news articles. Their results showed that their method can discover more useful emotion words, and their corresponding intensity improves their classification performance. Their method outperformed the (PMI)-based expansion methods as they consider both co-occurrence strength and contextual distribution, thus acquiring more useful emotion words and fewer noisy words.

2) *Chi-square (χ^2)*

Let n be the total number of documents in the collection, $p_i(w)$ be the conditional probability of class i for documents which contain w, P_i be the global fraction of documents containing the class i, and $F(w)$ be the global fraction of documents which contain the word w. Therefore, the χ^2-statistic of the word between word w and class i is defined as:

$$\chi i^2=n\cdot F(w)^2\cdot(pi(w)-Pi)^2F(w)\cdot(1-F(w))\cdot Pi\cdot(1-Pi) \qquad (2)$$

χ^2 and PMI are two different ways of measuring the correlation between terms and categories. χ^2 is better than PMI as it is a normalized value; therefore, these values are more comparable across terms in the same category.

χ^2 is used in many applications; one of them is the contextual advertising as presented by Fan and Chang. They discovered bloggers' immediate personal interests in order to improve online contextual advertising. They worked on real ads and actual blog pages from ebay.com, wikipedia.com and epinions.com. They used SVM (illustrated with details in the next section) for classification and χ^2 for FS. Their results showed that their method could effectively identify those ads that are positively-correlated with a blogger's personal interests.

Hagenau and Liebmann used feedback features by employing market feedback as part of their feature selection process regarding stock market data. Then, they used them with χ^2 and Bi-Normal Separation (BNS). They showed that a robust feature selection allows lifting classification accuracies significantly when combined with complex feature types. Their approach allows selecting semantically relevant features and reduces the problem of over-fitting when applying a machine learning approach. They used SVM as a classifier. Their results showed that the combination of advanced feature extraction methods and their feedback-based feature selection increases classification accuracy and allows improved sentiment analytics. This is because their approach allows reducing the number of less-explanatory features, i.e. *noise*, and limits negative effects of over-fitting when applying machine learning approaches to classify text messages.

3) Latent Semantic Indexing (LSI)

Feature selection methods attempt to reduce the dimensionality of the data by picking from the original set of attributes. Feature transformation methods create a smaller set of features as a function of the original set of features. LSI is one of the famous feature transformation methods. LSI method transforms the text space to a new axis system which is a linear combination of the original word features. Principal Component Analysis techniques (PCA) are used to achieve this goal . It determines the axis-system which retains the greatest level of information about the variations in the underlying attribute values. The main disadvantage of LSI is that it is an unsupervised technique which is blind to the underlying class-distribution. Therefore, the features found by LSI are not necessarily the directions along which the class-distribution of the underlying documents can be best separated.

3. Challenges of sentiment analysis

There are different challenges in sentiment analysis which is describe below. Implicit Sentiment and Sarcasm There is a chance that a sentence may contain implicit sentiment even though it is not having any word that earns sentiments. For an example, two statements are taken; "One should question that the stability of mind of the writer who wrote this book". In this above sentences don't have negative sentiment bearing words and no negative words are seen, although both are negative sentences. Thus identifying sentiment is important in Sentiment Analysis than syntax detection.

4. Applications of sentiment analysis

Domain Dependency: In this type of challenge words polarity changes from one domain to another domain in the domain dependency. For example, two statements; "The story was unpredictable." and "The steering of car is unpredictable." In first statement, in Sentiment express that is positive whereas the second statement express sentiment is negative.

Language Problem: In Opinion Mining English language is mostly used because of its resources availability means lexicons, dictionaries and corpora but User get attracted by using Opinion mining with language other than English like Hindi, French, Chinese, and Other languages

There different application of Sentiment Analysis. Sentiment analysis used in the movie review, product review, politics, public sentiment and social sites useful for people's opinion. Shown in the table there are various application of sentiment analysis in movie review by this user can get information about movie is good or bad or average by their star scale rating if movie is five stare we can predict that movie will be good if three star it will average review of movie. From the product review user can identify that product is good, excellent, average, and poor with the public opinion by their rating. When user have to settle on a choice user need to know others opinion. In the organization and associations dependably need to discover shopper or general assessments about their items and administration

Application of SA Different application different rating movie review product review politics public sentiment social sites mostly in the business or any affiliation required open or purchaser felling, it directed review, opinion polls, and focus on groups. The explosive growth of social media for example Twitter, Facebook, remarks, furthermore, posting in informal community destinations on the Web. Overviews, online journals, web journals Sentiment analysis applications have spread to each conceivable space, items, administrations, human services, and budgetary administrations to get-togethers and political decisions. These applications gave inspirations to inquire about in conclusion examination.

5. Conclusion and future work

This survey paper presented an overview on the recent updates in SA algorithms and applications. Fifty-four of the recently published and cited articles were categorized and summarized. These articles give contributions to many SA related fields that use SA techniques for various real-world applications. After analyzing these articles, it is clear that the enhancements of SC and FS algorithms are still an open field for research. Naïve Bayes and Support Vector Machines are the most frequently used ML algorithms for solving SC problem. They are considered a reference model where many proposed algorithms are compared to. The interest in languages other than English in this field is growing as there is still a lack of resources and researches concerning these languages. The most common lexicon source used is WordNet which exists in languages other than English. Building resources, used in SA tasks, is still needed for many natural languages.

Information from micro-blogs, blogs and forums as well as news source, is widely used in SA recently. This media information plays a great role in expressing people's feelings, or opinions about a certain topic or product. Using social network sites and micro-blogging sites as a source of data still needs deeper analysis. There are some benchmark data sets especially in reviews like IMDB which are used for algorithms evaluation.

In many applications, it is important to consider the context of the text and the user preferences. That is why we need to make more research on context-based SA. Using TL techniques, we can use related data to the domain in question as a training data. Using NLP tools to reinforce the SA process has attracted researchers recently and still needs some enhancements.

References

[1] Mikalai Tsytsarau, Themis PalpanasSurvey on mining subjective data on the web Data Min Knowl Discov, 24 (2012), pp. 478-514.

[2] Wilson T, Wiebe J, Hoffman P. Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of HLT/EMNLP; 2005.

[3] B. LiuSentiment analysis and opinion mining Synth Lect Human Lang Technol (2012)

[4] Adnan Duric, Fei Song, Feature selection for sentiment analysis based on content and syntax models Decis Support Syst, 53 (2012), pp. 704-711.

[5] Aggarwal Charu C, Zhai Cheng Xiang. Mining Text Data. Springer New York Dordrecht Heidelberg London: Springer Science Business Media, LLC'12; 2012.

[6] ThomasL. Griffiths, Mark Steyvers, David M. Blei, Joshua B. Tenenbaum Integrating topics and syntax Adv Neural Inform Process Syst (2005), pp. 537-544

[7] Walaa Meddhat, Ahmed Hassan, Hoda Korashy "Sentiment analysis algorithms and applications: A survey, Ain Sham University, Faculty of Engineering, Computer & Systems Department, Egypt 19 April 2014.

[8] Xing Fang and Justin Zhan, "sentiment analysis using product Review data" Department of computer science, North Carolina A&T State University Greensboro, NC, USA, 2015 Springer journal.

[9] Ebru Aydogan and M. Ali Akcayol "A Comprehensive Survey for Sentiment Analysis Tasks Using Machine Learning Techniques" Department of Computer Engineering Gazi University Ankara, Turkey 2016 IEEE

[10] Chandni, Nav Chandra, Sarishty Gupta, Renuka Pahade, "Sentiment Analysis and its Challenges" International Journal of Engineering Research & Technology (IJERT) 2015

[11] Vishal A. Kharde, and S. S. Sonawane "Sentiment Analysis of Twitter Data: A Survey of Techniques" International Journal of Computer Applications, April 2016.

[12] Yelena Mejova, Padmini Srinivasan. Exploring feature definition and selection for sentiment classifiers. In Proceedings of the fifth international AAAI conference on weblogs and social media; 2011.I.T. Jolliffee Principal component analysis Springer (2002).

[13] Ankush Sharma, Aakanksha, "A Comparative Study of Sentiments Analysis Using Rule Based and Support Vector Machine," International journal of Advanced Research in Computer and Communication Engineering, volume 3, March 2014.

[14] Chin-Shrng Yang, Hsiao-Ping Shih, A Rule-Based Approach for Effective Sentiment Analysis" PACIS 2012.

[15] Prern Chikersal, Soujanya Poria, and Erik Cambria, "SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning" June 5 2015.