An Overview of Sentiment Analysis: Concept, Techniques, and Challenges

** Mudita Nalawat*

Abstract

"Data is the new oil," a statement given by UK based mathematician, Clive Humby very accurately suits the current scenario of the World Wide Web. The unimaginable usage of online opinion sharing platforms is resulting in the generation of a large volume of opinion-rich data. This opinion rich data can be in the form of reviews given by people regarding any product, brand, service or something related to their experience and emotion about any topic. This data can be used by organizations to know what their customers really think, want, and need. Sentiment analysis is one of the technologies working behind this analysis. It is the process of mining text to extract subject information which can help organizations understand the social sentiments behind their product or service. This paper focuses on the aspects related to sentiment analysis. Various techniques that can be applied to sentiment analysis are discussed in this paper. This paper also focuses on how sentiment polarity of the data can be obtained using various sentiment classification techniques. Various applications and challenges related to sentiment analysis will also be explained in this paper.

Keywords: Corpus, opinion, polarity, sentiment

I. INTRODUCTION

With the evolution of internet-based applications over the last decade, there is a huge increase in the volume of opinion-rich web resources like blogs, review sites, discussion forums, various social media websites, wiki, etc. People from all over the world share their opinions on various topics through these platforms. The opinions present in these platforms in the form of information can be used by companies to predict their customer's preferences about a particular product or service.

Sentiment analysis (also called *opinion mining*) is a technology that can be used to analyze the textual content written by people on various online platforms. It refers to the contextual text mining technique which can be used to extract general information from the source data that can be beneficial for businesses to find out the social sentiments of their brands, products or services while keeping online conversation under observation. The massive amount of data present online can help companies know and understand their customers needs and demands in a more proficient manner, the utmost objective of sentiment orientation behind a text.

The most exciting field of compiler science and Artificial Intelligence, that is, Natural Language Processing (NLP), which largely deals with human-computer interactions uses sentiment analysis for processing and analyzing the data. The analyzed data from SA can be categorized as positive, negative, or neutral, depending upon the kind of words, phrases, and sentences used in the text. This categorization can help the companies know about their product or services' positive and negative features depending upon customer reviews.

The most common areas where SA is applied are: product review, Twitter data analytics, and comment analysis of people over different social networking sites. SA will help in finding out people's opinion, attitude, and emotions toward certain individuals, events or topics.

The rest of the paper is structured as follows. Section II is dedicated to literature review of research that has already been done in the field of sentiment analysis. Section III presents the various data sources used for extracting the corpus for sentiment analysis. Section IV explains the scope of the sentiment corpus. In section V the process of finding corpus has been explained. Section

Manuscript received June 18, 2019; revised July 5, 2019; accepted July 8, 2019. Date of publication August 5, 2019.

* M. Nalawat is a student at International School of Informatics and Management (IIIM), Sector 12, Mansarovar, Jaipur, Rajasthan, India - 302 001. (email: muditanalawat@gmail.com)

DOI: 10.17010/ijcs/2019/v4/i4/147302

VI describes the various techniques that can be used for sentiment classification. In section VII some of the most common applications that are using sentiment analysis have been discussed. Section VIII is about the challenges faced in sentiment analysis, that is, opinion mining. Finally, the conclusion of the paper is presented in section IX.

II. REVIEW OF LITERATURE

A. Sentiment Analysis in Social Media Platforms: The Contribution of Social Relationships

In [1], Fan, Ilk, and Zhang proposed a framework for sentiment analysis which is different from existing sentiment analysis approaches that solely rely on textual content of a sentence for sentiment identification. Researchers contended that the current sentiment analysis systems are not fully profitable for analyzing the content on social media because most of the people on social media use non-standard languages like abbreviations, misspellings, emoticons or other languages that give unclear results [1].

They conducted experiments to compare the proposed approach of incorporating social relationship in the sentiment analysis against the data set collected from Facebook [1].

In this paper, researchers proposed two types of social relationships, that is, user-topic and user-user relationship to capture the internal and external causes of user sentiments. Then they combined traditional textual analysis with social relationship to improve sentiment process. The proposed framework is based on attribute theory which explains a person's behavior using two categories of causes: internal attribution (for example, attitude or personality), and external attributions (for example, other people). They argued that a person's sentiment towards an entity is influenced either by personal causes or by environmental causes. In this framework, researchers have considered two types of social relationships, that is, user-topic relationship and user-user relationship [1].

In the overall framework, the researchers have used two data sources. First, they have analyzed the text in sentences where they have used clustering technique to identify topics for sentence and next they have built user-topic and user-user relationship based on social network and social media platforms. Finally, they have combined both results and used a machine learning based approach such as SVM (support vector machine) to predict sentiments.

Results predict that this framework has advanced the accuracy of sentiment analysis from 85% (existing approach) to 89% (user relationship framework).

B. Sentiment Analysis and Opinion Mining: A Survey

[2] is a research survey done by Vinodhini and Chandrasekaran. The paper presents a thorough survey which covers the techniques and methods applied in sentiment analysis and some major challenges appear in the field. In the paper, researchers focused on the work of English and Chinese in sentiment analysis. They have provided a brief about various data sources used for sentiment analysis (opinion mining). They have explained the sentiment classification of data based on three levels namely: (a) document level, (b) sentence level, and (c) attribute level. They explained the importance of machine learning for finding the opinion of the writer through mining text written by them and then reaching a specified result based on the analysis [2]. Machine learning techniques that are put into practice for the applicability of sentiment analysis are (i) Naïve Bayes, (ii) Maximum entropy, and (iii) Support vector (SVM) machine. They found that SVM exhibits the best performance for opinion mining or sentiment classification.

The research also highlighted the role of negation in sentiment analysis due to its effect on the polarity of any text [2]. Major applications discussed in the paper for sentiment analysis encompasses forums hotspot detection, online advertising, etc. Movie review and product review are the other two most common areas where sentiment analysis can be used effectively.

The research concludes that although sentiment analysis techniques are advancing fast, there are lots of problems which need to be resolved. The main is the negation expressions, handling such expressions needs expert algorithms.

C. Twitter as a Corpus for Sentiment Analysis and Opinion Mining

In [3], Pak and Paroubek proposed a technique that can be used to collect corpus (written text) for the purpose of opinion mining and sentiment analysis. They collected the sentiment data as corpus from the most popular platform of microblogging, that is, Twitter. Linguistic analysis was done on the collected corpus and from the results, the researchers proposed a sentiment classifier that can classify and determine the sentiments associated with the document as negative, positive or neutral [3].

The corpus for the research was collected from text messages of popular newspapers and magazines from Twitter. 44 Twitter accounts were queried to collect the training set of objective text. The researchers used Tree Tagger (Schmid, 1994) to tag the post in the corpus. The research explained the whole process of training the classifiers. A multinomial classifier known as Naïve Bayes classifier was used by the researchers in [3] for building the sentiment classifier. From the results, they concluded that their proposed classifier can manage to determine the positive, negative, and neutral sentiments pertaining to a document. Naïve Bayes classifier is the classifier which forms the basis of the proposed classifier. The classifier uses features like N-gram and POS-tags.

D. Sentiment Analysis: A comparative study on different approaches

In [4], the researchers compared distinct techniques that are used for the process of sentiment analysis by evaluating various methodologies. The paper focuses on different methods of sentiment analysis along with different levels of analyzing sentiments. Different approaches like machine learning, lexicon-based,and rule-based analysis are described in this paper. Further different machine learning methods are also described.

III. DATA SOURCES

The volume of data generated by people over the internet is boundless. People from all over the world share their opinions, thoughts, and views on various platforms of the internet. These platforms can be termed as the data sources from where we can find out people's opinions and sentiments about any topic by analyzing the context written by them.

A. Microblogging

The microblogging website acts as an online broadcast platform that provides people with the capability to share their views by writing content in a limited number of words. There are various microblogging sites that are famous among users like Twitter, Tumblr, and Facebook. Among these, Twitter is the most popular microblogging site through which users can share their opinions by writing status messages in the form of "Tweets". These tweets can be used as the data source for the SA process.

B. Blogs

The increased usage of the internet has resulted in an increase in the number of blogs. In these kinds of platforms, bloggers share their opinion about a particular topic. Many bloggers share their life's daily events and express their feelings in the form of blog writing. Blogs can also be used as a data source for SA.

1) Review Sites: While purchasing any product the customer always wants to know the opinion of others who have already purchased the same item. Various shopping sites like Amazon ask their customers for the review of the products. Sometimes the reviews are positive, sometimes they are negative. SA can be done by taking these reviews as the data sources.

2) Data Set: Large number of data sets are available online, data of movie review, product review, and many other types of data sets are present online which can be used for SA.

C. Scope

Depending upon the level upto which the process of sentiment analysis has to be done, there are different levels of scope for analyzing sentiments.

1) Document Level: When the sentiment analysis (SA) is to be done for finding the sentiment of document or paragraph, document sentiment analysis can be used. This type of SA classifies the document expressing negative, positive, or neutral sentiments[8].

2) Sentence Level: The process of sentence-level SA analyzes each and every sentence of the document and then determines whether the polarity concerning the sentence is negative, positive or neutral [8].

3) Aspect Level: The mechanism of aspect level of SA organizes the sentiment concerning the definitive features of the entities. In this process, firstly the entity selection process is done and then features are selected for those entities.

IV. PROCESS

Analyzing the sentiments of any textual content involves a series of steps. The sentiment polarity of any data can be obtained by going through different layers of procedures that have to be followed for finding sentiment polarity.

Fig. 1 depicts the overall process that has to be followed for finding the sentiment polarity of any textual content.

First, the data is collected from the data sources. The data can be from any source like blog data, Twitter data, Facebook data, product review data, movie review data etc. Depending upon the requirement of the business organization, data is extracted from a specific source. The next step involves the selection of sentences or data for SA.

After selecting the sentences or data, the next step is of tokenization. In this step, tokens are created for the selected data. As an example, if the sentence is “Yesterday I was not feeling well,” then by the process of tokenization it will be broken down into “'Yesterday', 'I', 'was', 'not', 'feeling', 'well' “. The next step followed by tokenization is the POS, that is, Part-of-Speech tagging step, also known as word-category disambiguation or grammatical tagging. This step involves categorization of lexical items having similar grammatical properties. It is stated that the words which are in the same part of speech exhibit similar behavior [7].

Sentiment identification is the next step in the process. In this step, the sentiment is being identified from the corpus having some opinionative word or phrases. Sentiments behind those corpuses are being identified in this step. After identifying the sentiments, the next step is of feature selection where different features are selected from the corpus. The features are selected on the basis of kind of data or corpus selected, and the type of review given by the people. The next step to follow after the feature selection is the sentiment classification step where classification technique is applied to the corpus. The technique is selected depending upon the need of the business. The techniques that can be used for sentiment classification are discussed in the next section of this paper. The technique is selected depending upon the need of the business. Results are generated by using sentiment analysis techniques. From the results, the SA process finds the sentiment polarity of the data.

This is the complete process that can be followed to find the sentiment polarity of data set.

V. CLASSIFICATION TECHNIQUES

Sentiment analysis or opinion mining plays a major role in any organization to make decisions about any product or service. For better decision making, the correct technique should be selected so that it can give results of high accuracy.

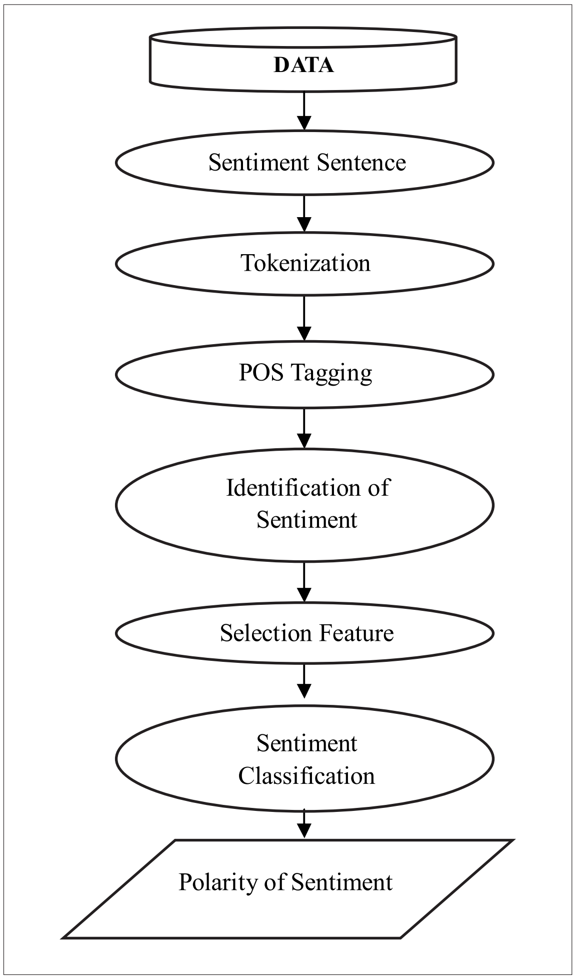

Fig. 1. Procedure to Find Sentiment Polarity

There are three types of techniques that can be applied for sentiment classification.

1) Sentiment analysis using machine learning approach.
2) Sentiment analysis using lexicon based approach.
3) Sentiment analysis using the hybrid approach.

The hybrid approach was introduced by combining the first two approaches, taking the best features of machine learning and lexicon based approach. This approach was introduced for gaining high accuracy, but in a broad way only the first two techniques are explained in this paper.

A. Machine Learning Approach

The machine learning approach is one of the most widely used approaches in sentiment analysis. The reason for using this approach is that it has higher accuracy. This

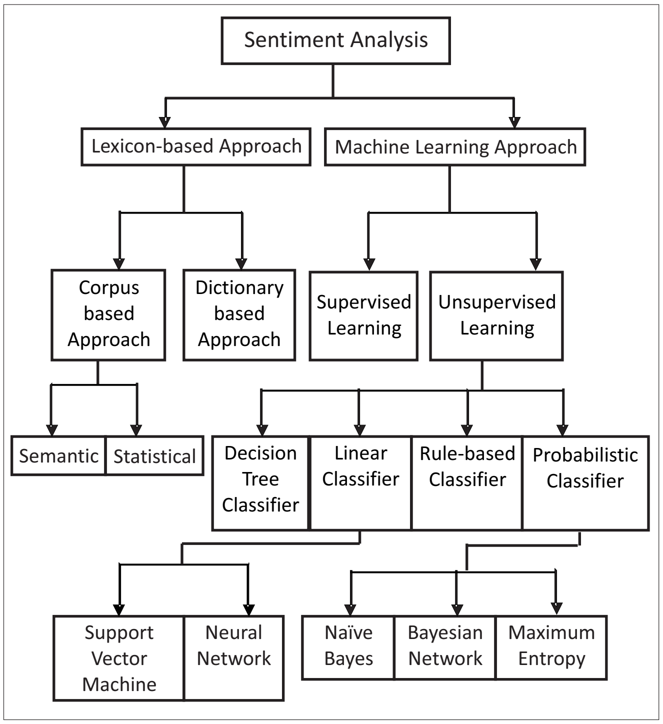

Fig. 2. Sentiment Classification Techniques

approach works by training the algorithm with some training dataset before applying it to the authentic data set [9]. In this approach, the algorithm is first trained by giving some inputs with known outputs to check whether it will work accurately or not with unknown data.

A number of techniques can be used in the machine learning approach. These are as follows:

1) Support Vector Machine: This technique requires a large number of training data sets for making the sentiment classification possible. It is a non-probabilistic classification technique. In this method, the data is divided by applying some decision boundaries to the data. On the basis of the features taken, some data will lie inside the boundary and some will lie outside the boundary. By analyzing the data on either side of the boundary, sentiment classification is done.

2) N-gram Sentiment Analysis: N-gram method is used in probability and linguistics, which is treated as a continuous progression of '*n*' number of items from the chain of text given. The '*n*' items can probably be words, letters, syllables, phonemes, or base pairs. They are selected according to the type of application. The *n* in the n-gram can be any number like 1-gram, 2-gram, etc. N-gram model utilizes four different kinds of lexicons specifically, sentiment phrase lexicons, sentiment strength lexicons, lexicon with aspect, and exception lexicons.

3) Naïve Bayes Method: It is a kind of probabilistic classifier method that is mostly used when the number of training sets is less. This classifier works on the principle of Bayes theorem which is also a probabilistic classifier [7]. This classification assumes that all the features that approach as parameters in the data set, and the machines are independent of each other.

4) Maximum Entropy Method: Maximum Entropy method, also known as the conditional exponential classifier is somehow related to naïve Bayes classifier. The difference is that it allows features not to be independent of each other. In Maximum Entropy classification, the classifier is parameterized by a series of weights that are utilized for combining the joint-features and brings out from them the set of features through encoding. This encoding is used for mapping each pair of features that seem to be the label to a vector. For the distribution to be uniform, the uncertainty in this technique is maximum. A term known as entropy can be used to measure the uncertainty.

5) K-Nearest Neighbor Method: The K-Nearest Neighbor (K-NN) method is a non-parametric method used for the sentiment classification process. This method is based on the certainty that the classification of an instance will be somehow like those which are nearby in the vector space. In this method, a training set and unknown sample are given, and the distances between the unknown sample of data and all the sample data in the training dataset are computed. The distance with the smallest value coincides with the sample associated with the training dataset nearest to the unknown sample.

6) Multilingual Sentiment Analysis: Multilingual sentimental analysis is the method in which text written in different languages can be analyzed for SA in the same way as it processes in English. People from different language backgrounds share their opinions and views in different languages, so this technique can be used in such kind of text SA. In this method, first the language is identified using language model and after identification, it is converted into English using standard translation software.

7) Lexicon Based Approach: Lexicon based approach depends on sentiment lexicons, which are the collection of familiar and precompiled sentiment terms. No training data is needed for the algorithm in this approach of sentiment analysis. In this approach, dictionaries of words are annotated with their semantic polarity and sentiment strength. The output is then used to calculate a score for the polarity of the document[7]. This method requires powerful linguistic resources for analyzing the contextual data. One of the best features of this approach is that it does not require any specific procedure for learning and the labeled data

VI. APPLICATIONS

The applications of sentiment analysis are increasing day by day with the advancement in the features of sentiment analysis. There are various applications of sentiment analysis in the current market, and will surely increase in number in the coming years.

Some of the areas where SA is applied to a great extent are as follows:

A. Product Review

A large number of customers now a days use online platform for purchasing. Everyone wants to know about the review of items that they want to purchase before actually purchasing them. Reviews by people who have already purchased an item can help other people in the purchase decision. These reviews also help the product manufacturer to know the preference of their customer and to know about the upgrades that have to be done for their products [6].

B. Movie Review

Sentiment analysis can be applied to movie review data to find out the chances of the success of any movie. Reviews about the actor, actress, director, producer, etc. are collected and then analyzed. This analysis can help the movie makers to find the chances of their upcoming movie to be a hit, flop or moderate.

C. Trends

Upcoming trend in the market can be found out with the help of SA. In this, various platforms of social media are analyzed to find out what people are preferring. Depending upon the features with greater positive reviews, trends are analyzed [5].

D. Politics

The time when politicians want to know about the voters' view, sentiment analysis plays a major role in identification. In politics, sentiment analysis can also be used for predicting the result of polling by analyzing the opinions and sentiments of people.

VII. CHALLENGES

Even though sentiment analysis technology has reached a level where the task of analyzing any text for finding sentiment or opinion behind it has become much simpler as compared to older times, still there are various challenges faced by the analyst in sentiment analysis. These challenges begin to be a barrier in analyzing the correct meaning of sentiment and in identifying satisfactory sentiment polarity.

The most common challenges faced by the analysts during sentiment analysis are:

A. Data Scarcity

Data is present everywhere over the internet and the problem is that this data is very scattered which creates difficulty in analysis. The challenge in sentiment analysis is to bring all the scattered data together in such a way that it allows for accurate analysis and manipulation.

B. Multilingual Sentiment Analysis

Even though there are techniques that can be used in multilingual SA. In such a technique, first the text written in a language other than English is converted into English with the help of translators and then SA is performed on the translated text. Still, there are some areas where multilingual SA lacks. As an example, while translating the data into English, many researchers have found that there is a loss in the degree of sentiment after conversion as compared to before conversion. This is the challenge faced by many researchers during multilingual sentiment analysis.

C. Sarcasm Detection

One of the major challenges faced during sentiment analysis is sarcasm detection. Many a times, people convey their message in a way which totally conflicts the context. These messages create uncertainty in detecting the correct sentiment polarity of any data. When people share their opinion regarding any topic in a sarcastic way, the sentiment classifier interprets the data as if it is a genuine opinion of the writer which creates unambiguity in the results. There is a need for research that could be dedicated to the technique that could accurately differentiate between genuine content and sarcastic content.

D. Fake Review

Many people sometimes give fake reviews about any product or service which cannot be determined by the analyzer, whether it is a real review or a fake review. Such kind of review generates incorrect results in the overall SA. This challenge is a major one in the field of SA [10].

VIII. CONCLUSION

Sentiment analysis has proven to be one of the emerging fields of data analytics. The large volume of opinion-rich data available at different data sources has resulted in better prediction about the needs and demand of people for business organizations. Advancement in NLP (Natural Language Processing) and Machine learning has proven the fact that companies can get deepest information about their customers if they thoroughly analyze the content and can get maximum of profit from the obtained results. This paper examined the overall process of finding the sentiment polarity of any textual content. Various techniques were discussed in the paper that can be applied to the corpus for sentiment classification. By analyzing the techniques, it has been found that different type of corpus works on different techniques but it has also been found that support vector machine (SVM) outperforms other techniques in many ways.

However, even though much work has been done in the field of sentiment analysis, there is a requirement of more work to be done for improving the performance. It has been found that there are a lot of problems in the field of sentiment analysis that remained unsolved. There are various challenges in the field of study that need to be tackled. There is still a need for enhancement in the field that could be dedicated to these challenges.

REFERENCES

[1] S. Fan, N. Ilk and K. Zhang, “Sentiment analysis in social media platforms: The contribution of social relationships,” *pp. 1-9, 2015.*[Online]. Available: https://pdfs.semanticscholar.org/6c71/3a2fe4b619a699f0827faf232983fc35ef5d.pdf

[2] G. Vinodhini and R. M. Chandrasekaran, “Sentiment analysis and opinion mining: A survey,” *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. *2*, no. *6*, 2012. [Online]. Available: https://pdfs.semanticscholar.org/261e/26ae134b8f63270dbcacf2d07fa700fdf593.pdf

[3] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” In *Proceedings of the Seventh conference on International Language Resources and Evaluation* (LREC'10), pp. 1320-1326.[Online].Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

[4] M. D. Devika, C. Sunitha and A. Ganesh, “Sentiment analysis: A comparative study on different approaches,” *Procedia Computer Science, 87*, pp. 44-49,2016. Doi: https://doi.org/10.1016/j.procs.2016.05.124

[5] W. Medhat, A. Hassan and H. Korashy, “Sentiment analysis algorithm and applications: A survey”, *Ain Shams Engineering Journal,* vol. *5*, no. *4*, pp. 1093- 1113, 2014. Doi: https://doi.org/10.1016/j.asej.2014.04.011

[6] X. Fang and J. Zhan, “Sentiment analysis using product review data,” *Journal of Big Data,* vol. *2*, no. *1*, 2015. Doi: https://doi.org/10.1186/s40537-015-0015-2

[7] H. Rahmat. P. and T. Ahmad, “Sentiment analysis techniques: A comparative study,”*International Journal of Computational Engineering & Management*, vol. *17*, no. *4*, pp. 25-29, 2014. [Online]. Available: https://pdfs.semanticscholar.org/a2cb/b062f5a254866bb8fdff4f814130bfcc835f.pdf

[8] B. Liu, *Sentiment analysis and opinion miningSynthesis Lectures on Human Language Technologies,* 2012. USA: Morgan and Claypool Publishers.

[9] S. M. Vohra and J. B. Teraiya, “A comparative study of sentiment analysis techniques,” *Journal of Information Knowledge and Research in Computer Engineering*, vol. *2*, no. *2*, 2013.

[10] D. M. Hussein, "A survey on sentiment analysis challenges,"*Journal of King Saud University-Engineering Sciences,* vol. *30*, no. *4*, pp. 330-338, 2018. Doi: https://doi.org/10.1016/j.jksues.2016.04.002

About the Author

Mudita Nalawat completed Masters of Computer Application (MCA) from International School of Informatics and Management, Jaipur in the year 2019. She completed her graduation degree in computer application (BCA) from University Maharani's College, Jaipur, Rajasthan. She has always been interested in browsing new trends and technologies that are booming in the field of computer science.