A Study on Sentiment Analysis on Social Media Data

K. N. Manasa and M. C. Padma

Abstract Sentiment analysis which is otherwise also called sentiment mining or opinion mining is the process of ascertaining and categorizing the positive, negative, or neutral opinion of the speaker or writer about a specific product, service, etc., in essence. The development of the user-generated content in social media has opened new prospects to extract knowledge from the opinions. Sentiment analysis categorizes the opinion of a writer into positive, negative, and neutral. This paper presents a detailed study on sentiment analysis process, various tools used for data collection, importance and the sources of user data. The importance of machine learning in sentiment analysis with various preprocessing steps and the ontology for sentiment is also discussed.

Keywords Sentimental analysis · Machine learning · Opinion mining · Ontology · Data lake

1 Introduction

Human beings by nature have exhibited the behavior of articulating their feelings, emotions, and opinion on every aspect of life. The invention of social media platform has opened new doors of freedom for people to express their opinions and views. The Web and social media have a direct impact on every aspect of peoples' lives. With the rise of enormous social media platforms, people get an opportunity to share their thoughts on this platform. Various social media sites are available to express user's feelings [1].

Most of the times opinions expressed are considered to be subjective. Manual classification of these data into different classes of opinions is a challenging task [2]. Natural language processing was one of the earliest methods which tried to meet this

K. N. Manasa (✉) · M. C. Padma
PES College of Engineering (Affiliated to University of Mysore, Mysuru),
Mandya 571401, Karnataka, India
e-mail: manasaphd17@gmail.com

V. Sridhar et al. (eds.), *Emerging Research in Electronics, Computer Science and Technology*, Lecture Notes in Electrical Engineering 545,
https://doi.org/10.1007/978-981-13-5802-9_58

great challenge. But still it remains as an open area to be explored more [2]. A natural language processing task that helps in the identification and extraction of information is known as sentiment analysis (SA). SA is a process of categorizing the attitude, opinion, or emotions of the person or the customer from a given piece of writing. SA also called opinion mining helps to identify the opinion of customers or followers toward a product or an event. This process initially involves conversion of text into numerical data. It involves an approach to classify the text into positive, negative, or neutral category. The main challenge in the social media data is identifying and classifying the human opinions present in conversations which have various meaning and nuances. Various algorithms and technologies are being used to collect, analyze, and identify the opinions present in the text [3].

2 Importance of Social Media Data and Its Sources

Social media data are the huge and rich source of human behavior that provides a way to understand people and their emotions. It has created an abundance of information. Individuals do various activities on social media like write comments and reviews on products or events, tweets, retweets, post, share, and discuss consistently. With this expanded movement, social media plays an important role and provides new opportunities to analyze human behavior and information dissemination at a scale that is generally difficult to perform. Social media data are the major source of text mining and predictive analysis. Literature survey reveals that several attempts have been made to analyze social media data for predictions like stock market, rainfall, and sales. [3]. Normally, huge amount of statistics is produced from social networks, blogs, and other media such as Facebook, Twitter, Linkedin, and WhatsApp. This massive record consists of very important opinion related information that can be used to benefit agencies and industries. Manual tracking and extraction of these useful records are not always possible. In this context, SA plays an important role [3].

Generally, it is easy to understand sentiment(s) of one individual, but understanding sentiments of a gathering is a major challenge. Several techniques such as data mining, text mining, and machine learning are used for this purpose [4].

3 Machine Learning in Sentiment Analysis

Almost all Machine learning algorithms for sentiment analysis follow standard steps to classify the sentiments as explained below [5]:

Noise Removal: Cleaning the data to extract relevant data from irrelevant data which increases the ability of an algorithm to predict based on the training data set.

Classification: Categorizing the data into positive and negative class.

Named Entity Recognition: In order to predict actual meaning of the comments, say, in social media such as Facebook, it is obvious to extract entities like sender, receiver, and the aspects of their conversation and then classify them as positive or negative.

Subjectivity Classification: Sentences can be classified into subjective or objective. Subjectives are expressions on any attributes, events with the properties. Objectives are opinions that describe feelings toward entities.

Feature Selection: The features can be unigrams, bigrams, or n-grams with or without punctuations.

Sentiment Extraction: It can be done using unsupervised and supervised machine learning algorithms such as extra trees classifier, random forest classifier, naïve Bayes, and support vector machine.

4 Data Collection and Storage

In order to perform SA on social media data, various tools and technologies available for data collection, storage, and access are utilized. The main challenge here is handling huge volume of data at minimum managing costs [6]. In general, big data comprises structured and unstructured data. Social media data is unstructured since it does not have a predefined information model to portray its contents. Data warehouse generally handles structured data but not unstructured data [7]. With enormous growth in unstructured data in the present days, an effective technology to handle them is definitely needed. Data lake which stores data irrespective of its format meets this demand. It can collect single or multiple type data of hundreds of terabytes, manage to store them in their native format, and help to perform data analytics from various sources on the data. This concept is mainly used in machine learning, analytics, and visualization. Azure Data Lake tools Plug-in can be used to perform SA on social media data [8].

5 Ontology for Sentiment Analysis

To perform SA on social media data, an appropriate ontology-based model is required which describes various entities and their properties. Ontology can be created using various methods to identify objects, attributes, and their relationship for the domain. Ontology-based SA model consists of classes, subclasses, objects, and their properties. The objects are stored in the database, and ontology model can be queried using query language like SPARQL to retrieve the data. Using ontology, a dynamic model can be developed that can predict objects regardless of the domain being used. Features of the domain are extracted by building ontology which helps in getting the refined SA. Feature-based sentiment analysis gives the best result while analyzing the sentiment [9].

6 Related Work

SA can be done on any data collected from social media networks. Most of the researchers prefer to collect data from Twitter since it has data from all the fields. Several works have been done on Twitter SA using machine learning techniques. Ren and Wu [4] have done an experiment to predict the unknown user-topic opinions using lexicon-based approach. An accuracy of 0.6046-F1 score has been obtained as a result. A machine learning naive Bayes with rapid miner software and R studio has been used to classify the sentiments of Twitter data by Das et al. [10], and out of 1298 tweets, they have classified 395 negative tweets, 187 neutral tweets, and 716 positive comments. A text analysis has been done by Jain and Dandannavar [5] for Twitter data using machine learning algorithms such as naïve Bayes and decision trees. The limitation of their work is that as the lexicon size increases, it gives an error and consumes more time. A support vector machine, naive Bayes, and maximum entropy algorithms have used by Neethu and Rajasree [11] for sentiment classification using Tweets on electronic products analyze the outcome of domain information. An accuracy of 89.5% has been obtained. Hasan et al. [12] have explained about SA using twitter data and have classified the data collected as sentiments based on a method of contextual phrase-level polarity. But they have not considered patterns of political parties based on Twitter reviews. Bravo et al. [13] have used microblogging messages for sentiment classification using a novel approach with the help of collection of numerous available lexical resources. They have obtained more than 5% accuracy and F1 points than the existing methods. Since the analysis and processing are difficult with huge data, the Sehgal and Agarwal [14] have used Hadoop to make the process simple. They have considered movie review from Twitter and have obtained an accuracy of 72.22%. Walha et al. [15] have proposed a lexicon opinion analysis approach for extracting the extracts from the sentiment polarity of informal text from social media data particularly Twitter data.

Many research works have been done on ontology based for SA which retrieves data from social media sites, and it helps to find words which are not available on the corpus by finding the nearest mean of the words. To analyze the negative emotions of the customer from Twitter, Takor and Sasi [9] have proposed an ontology-based sentiment analysis process. More works need to be done to assign the properties with the object automatically. The result obtained from their method is 54 negative comments out of 250 tweets. Haider [16] has proposed an ontology model for feature-based SA to analyze the customer review on normal mobile phones using smartphones. High accuracy has been obtained compared to existing model. Drawback of the work is that manual data collection has been done which is time-consuming. An ontology model with user query, processing was developed by Sam and Chatwin [17] in order to understand the responses obtained in market from online consumers. The result obtained was 90% with emotional tolerance index zero. Takor and Sasi [9] have proposed an ontology-based model to retrieve and analyze the customers' tweet with negative sentiments including its intensity, and a rule-based engine is developed for the same to apply in real-world applications. Song et al. [6] have investigated online diffusion of information that creates fear with regard to infection of a disease in the Middle East through SA of social media data.

Finding the intensity level of sentiment is another area of SA. The strength of the sentiment can be determined by the polarity or the scope of sentiments. Tomar and Sharma [18] have developed an analysis strategy for detecting text message polarity using SentiWordNet. The results obtained for the movie reviews are 73.2% positive, 72.1% negative, and 66% neutral. The outcomes from the hotel reviews for positive, negative, and neutral comments are 73%, 69.8%, and 60% respectively. Thelwall et al. have concentrated on finding out the weighage of each sentiment term. They have extracted emotions from nonstandard spelling in the reviews and also have corrected misspellings using machine learning approach. Gatti and Guerini [19] have addressed the problem of computing the intensity of each word in prior by beginning from the polarity of each word. As a result, it gives the index value for the polarity strength in the range of -1 and 1. Various languages have been considered for the work, and the result obtained was mean—0.708, precision—0.707, F1 score—0.707. Song et al. [20] have found out the weight of positive and negative words and also have suggested that the proposed work could give better accuracy when compared to existing two algorithms such as multivariate Bernoulli naïve Bayes and multinomial naïve Bayes (MNB).

7 Existing Techniques for Sentiment Analysis

The existing work on sentiment analysis can be classified into different points of views: technique used, view of the text, level of detail of text analysis, rating level, etc. From a technical point of view, there are four basic approaches such as machine learning, lexicon-based, statistical, and rule-based approaches. The machine learning approach is used widely because of its ability to handle automation and to deal with large and small data sets. The lexicon-based approach is used to calculate polarity of sentiments from the reviewed data for getting lexical option based on dictionary approach or corpus approach. Semantic orientation is a measure of subjectivity and opinion in text. The rule-based approach classifies based on the number of positive and negative words. This approach uses different rules for classification such as dictionary polarity, negation words, booster words, idioms, emoticons, and mixed opinions. Most of the existing techniques of classification techniques classify the data inaccurately, and most of the sentimental analysis is done on English text due to the lack of dictionary and corpus in other languages.

8 Research Challenges in the Domain of Sentimental Analysis

SA is a vast research area with lots of potential research problems that can be addressed to add value to the decision-making process. As most of the analysis is on classification of sentiments, the analysis of opinion's strength will boost the classification of social media data very efficiently, where a combination of multiple sentiment dimensions can be proposed. An understanding of identification of sarcasm and irony in sentences combined with other features is also an aspect that affects the identification of intensity of opinions.

Embedding of opinions in ontology would help in answering intelligent opinions and queries which has not been much researched so far. A semantic framework using deep learning would help in reducing the entity relationship to add value to the area of opinion mining. A combined approach of rule-based learning along with machine learning techniques can give better results.

Though good and impressive results have been shown in this area, it still requires much research as most of the decision-making, business sales, and forecasting models are based on the opinion from social media data.

9 Conclusion

SA is a major research area which mainly focuses on understanding the behavior and the expectations of society in the form of writing in the social media sites. This is an easy and cost-effective method of identifying how people feel about a specific topic or a product. Twitter is considered as one of the popular social media sites for data collection since it is a platform which allows users to discuss on various topics, and through API, the complete data can be accessed. This paper gives an insight into the recent updates and innovations in SA. A majority of the recent works done are based on machine learning method. As most of the business decisions are based on opinion analysis, an ontology-based machine learning approach can be explored for making decisions better.

References

1. Dijck JV (2013) The culture of connectivity a critical history of social media. Oxford University Press, New York
2. Chaturvedi I, Cambria E, Welsch R, Herrera F (2018) Distinguishing between facts and opinions for sentiment analysis: survey and challenges. Inf Fusion 44:65–77
3. Sun A, Lachanski M, Fabozzi FJ (2016) Trade the tweet: social media text mining and sparse matrix factorization for stock market prediction. Int Rev Financ Anal 48:272–281
4. Ren F, Wu Y (2013) Predicting user-topic opinions in twitter with social and topical context. IEEE Trans Affect Comput 4(4):412–424

5. Jain AP, Dandannavar P (2016) Application of machine learning techniques to sentiment analysis. In: 2016 2nd international conference on applied and theoretical computing and communication technology (iCATccT). IEEE, Bangalore, India, pp 628–632
6. Song J, Song TM, Seo DC, Jin DL, Kim JS (2017) Social big data analysis of information spread and perceived infection risk during the 2015 Middle East respiratory syndrome outbreak in South Korea. Cyberpsychol Behav Soc Netw 20(1):22–29
7. Walker C, Alrehamy H (2015) Personal data lake with data gravity pull. In: 2015 IEEE Fifth international conference on big data and cloud computing (BDCloud). IEEE, Chennai, India, pp 160–167
8. https://msdn.microsoft.com/en-us/azure/data-lake-analytics/u-sql/sentiment-analysis-u-sql
9. Thakor P, Sasi S (2015) Ontology-based sentiment analysis process for social media content. Procedia Comput Sci 53:199–207
10. Das DD, Sharma S, Natani S, Khare N, Singh B (2017) Sentimental analysis for airline twitter data. IOP Conf Ser Mater Sci Eng 263:042067
11. Neethu MS, Rajasree R (2013) Sentiment analysis in twitter using machine learning techniques. In: 2013 fourth international conference on computing communications and networking technologies (ICCCNT). IEEE, Tiruchengode, India, pp 1–5
12. Hasan A, Moin S, Karim S, Shamshirband S (2018) Machine learning-based sentiment analysis for twitter accounts. Math Comput Appl 23(1):11
13. Bravo-Marquez F, Mendoza M, Poblete B (2013) Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In: Proceedings of the second international workshop on issues of sentiment discovery and opinion mining. ACM, Chicago, USA
14. Sehgal D, Agarwal AK (2018) Real-time sentiment analysis of big data applications using twitter data with Hadoop framework. In: Soft computing: theories and applications. Springer, New York, USA, pp 765–772
15. Walha A, Ghozzi F, Gargouri FA (2016) Lexicon approach to multidimensional analysis of tweets opinion. In: 2016 IEEE/ACS 13th international conference of computer systems and applications (AICCSA), pp 1–8
16. Haider SZ (2012) Ontology-based sentiment analysis case study (Master degree project). University of Skovde, pp 5–67
17. Sam KM, Chatwin CR (2013) Ontology-based sentiment analysis model of customer reviews for electronic products. Int J e-Bus e-Manag e-Learn 3(6):477–482
18. Tomar DS, Sharma P (2016) A text polarity analysis using Sentiwordnet based an algorithm. Int J Comput Sci Inf Technol 7(1):190–193
19. Gatti L, Guerini M (2012) Assessing sentiment strength in words prior polarities. In: Proceedings of the 24th international conference on computational linguistics. COLING ’12, Mumbai, India, pp 361–70
20. Song J, Kim KT, Lee B, Kim S, Youn HY (2017) A novel classification approach based on Naïve Bayes for Twitter sentiment analysis. KSII Trans Internet Inf Syst 11(6):2996–3011