The Sentimental Analysis of Social Media Data: A Survey

Vartika Bhadana and Hitendra Garg

Abstract Nowadays, machine learning plays a very important role in every field. For recommendation systems, user feedback is relevant because they contain different forms of emotional details that may affect the reliability or consistency of the recommendation. Online reviews, comments are very helpful in selecting the said items and services as it gives real feedback about the quality of these items and services. The categorizations of these items based on feedback provided by actual users are known as sentimental analysis. In this study, we described various machine learning techniques and parameters used for the sentimental analysis of reviews, comments, and feedback available on health care, Facebook, Twitter, and other social media networking sites. The study reveals that the most commonly used approaches are machine learning and deep learning.

Keywords Machine learning techniques · Lexicon-based approaches · Corpus-based approaches · Deep learning · Neural networks

1 Introduction

As sentiment analysis is a major problem in today's world, people are facing difficulty in predicting things. As people, we still seem to attract like-minded people. Also, surveys show that they are not confident socializing with individuals with common values, what people should believe, and who will help us accomplish those goals. Etymologically, individuals prefer to connect to similar-minded groups. Many clusters make up a group. Modularity is one of the key factors considered during the quantity calculation of populations. If the features of the clusters are studied in-depth, it may be helpful to define the unique character profile of particular clusters or groups of individuals with like-minds. As sentiment analysis can be in many forms,

V. Bhadana · H. Garg (✉)
GLA University, Mathura, UP, India

V. Bhadana
e-mail: vartika.bhadana_mtcs19@gla.ac.in

P. K. Singh et al. (eds.), *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, Lecture Notes in Networks and Systems 203,
https://doi.org/10.1007/978-981-16-0733-2_17

it can be in health care, social media, e-commerce Website, and language prediction. Machine learning techniques are very helpful in predicting sentiments. As social media has many reviews, individuals cannot predict that the reviews or comment is of sentiment good, bad, excellent, or average. Some reviews are in different languages which cannot be understood by an individual. If citizens have to book the hotel for accommodation, there also sentiments are major issues that time, also it is difficult to predict which hotel is best, and many reviews can be fake also which are best for that particular hotel. So, to overcome these problems, some techniques are been used which will predict these issues from the different datasets of different types.

The proposed survey paper aims to give knowledge about the sentiment in different fields. Sentiment analysis uses different techniques to find the best accuracy among the sentiments.

2 Related Work

These sentiment analyses have been discussed under the following headings.

2.1 Sentiments Polarity Improvement

As the sentiment detection polarity describes that there are so many long reviews or comments that can be computed in the short meaningful sentences, firstly, they follow five policies like.

2.1.1 Most Occurring First (MOF)

This approach defines the most occurring goal as the overall review's main aim and then measures the review's polarity dependent on the review's main objective. This technique is analogous to the voting process, which is effective when the analysis has a dominant purpose. Nevertheless, in the study, two or three targets may be set for the same frequencies. The following methods should be used instead, in these situations.

2.1.2 Most General First (MGF)

Under this method, the most general objective is known as the main objective of the overall analysis, and therefore, the analysis polarity is determined based on the main objective. An ontology-based approach is used to identify the most general goal.

2.1.3 Most Specific First (MSF)

This approach defines the most important target as the overall review's main target and then measures the review's polarity based on the review's main objective.

2.1.4 First Occurring First (FOF)

It takes the first target listed in the analysis as the overall review's key target and then determines the review's polarity based on that goal.

2.1.5 Last Occurring First (LOF)

This approach considers the last goal found in the analysis as the key objective of the whole study and then measures the polarity of the review based on this objective.

The next step is the part of speech (POS) tags in which reviews are been compressed and the lexicon approach is been applied to it. Basiri et al. focused on the Persian language and have used the dataset of hotels and movies. As the Persian language has some drawback also as the lexicon are not precise as they are for the English language. There is no large dataset for the Persian language to train and test for ML techniques that will affect the polarity. The Persian language can have grammatical mistakes as compared to English. As they choose for the lexicon approach because it is simpler than the ML techniques as lexicon does not require training data, it is domain-independent.

Ontology is a systematic classification of definitions and the interaction between them. That may be as basic as definition taxonomy, which may include axioms and constraints to describe multiple facets of the natural world. The former is commonly considered a lightweight ontology, while the latter is regarded as an ontology of high weight. Chi et al. developed a lightweight ontology from the ground up, as there was no ontology existing for Persian principles relating to their datasets [1].

The system proposed typically has four major steps. The first section, pre-processing, involves phases of tokenization, standardization, and POS marking that are implemented sequentially on analysis sentences. The next step is called "extraction of possible words" and is responsible for separating future words from the statements. Typically, such words include adjectives, adverbs, and negations. After identifying the goals of each paragraph, the third stage called "target recognition" is to determine the key target for the whole study. The final stage is the classification process, which is responsible for determining the review's final polarity [2].

In the Persian language, certain pre-processing measures are identical to the Arabic language. For example, in both languages, the tokenization stage is more difficult than in English since they do not use capital letters and do not have specific punctuation rules. Another example is the stemming method in the Persian language, like the Arabic language, where in addition to adding prefix and postfix to the stems to produce a new grammatical form, it is also possible to attach infixes to the stem

which makes the stemming method more complex than English in the Arabic and Persian languages [2]. Rathi et al., info on the mining of thoughts, how the importance of polarity deals with positive and negative and how to cope with Roman language reviews and journals.

Two pre-processing steps are expected to be performed on the dataset prior to the main process: tokenization and normalization.

Boundaries of words are defined in the tokenization process which is necessary for the next steps. Use functional units like phrase, sentence, and separators like space and line, and there suggested method tokenizes reviews. Normalization is a typical stage in text processing in which the different representations of a single word are combined and translated into the same type.

Basiri et al. have taken the preview dataset which is a labelled dataset, so analysis has two labels in it: one to indicate its polarity and the other to define its principal goal. The target of reviews of unknown destinations is defined as "implied." For example, the target is implied in the analysis seen in Example 1.

Example 1 Hello, it looks great, but now you can purchase two Android phones with the same or better specifications without restriction when you update the device and more. Please dismiss mark prejudice. Something costlier cannot automatically be easier.

Best accuracy was measured by the lexicon approach and $F1$ measure, and among the five identification techniques, MGF scores the best rank [1].

A hybrid ML-based approach is been implemented to integrate SVM and Naïve Bayes in the online Persian film review data collection to identify consumer comments as either favourable or negative. As a result, an increase compared with previous studies has been identified. Hajmohammadi et al. studied the close association between the extraction process of the characteristic and the findings obtained [3].

2.2 Sentiments in the User Rating on Social Media Sites

It is another issue that describes the deep leaning technique. The proposed work is a profound learning model for processing user comments and creating a potential user rating for user suggestions. First, the program uses sentiment analysis as the input points to create a vector of the function. Next, the device implements dataset noise reduction to improve user rating classification. Finally, for the suggestions, a deep belief network and sentiment analysis (DBNSA) should achieve data learning. The experimental results suggested greater precision of this device than conventional methods. Analysis of the feelings is based on a lexicon of thought. An opinion lexicon is a word dictionary that communicates the polarity of words by positive or negative emotions such as happy, good, bad or disgusting. In sentiment analysis, these opinion terms are used as the primary predictor for measuring the user's opinions. Many

public lexical sets have become accessible in recent years, such as SentiWordNet. The first important task in sentiment analysis is to define the opinion objectives (aspects, persons, and issues of recognition of topics) on which opinions are expressed. Next, the lexicon of opinion has to be built.

The paper introduces three steps of noise reduction. The program extracts user comments in the first step that contain just ten words or fewer, and for which the user comments do not contain a word from the opinion lexicon. The second step is the identification of negative terms used in the lexicon of opinion. For example, "not bad" is a negative term that is included with "evil" in the positive opinion lexicon. Use the Glove algorithm to measure the sense of similarity for the negative term opinion lexicon [4]. In the third step, they delete all remarks which are categorized as good impressions, making the lexicon target the negative impression, and vice versa. Using a DBN and sentiment analysis, they propose a novel approach to predict user ratings from user comments. The sentiment analysis generates the feature vector based on the good and bad opinions of products or services as recorded by users in their comments. They also enforce noise reduction procedures that remove short comments, comments with no speech in them, and false rating comments. The DBNSA model outperforms other baseline models in the experimental section and outperforms baseline models in training failure performance, accuracy, and recall on Yelp and Amazon datasets [5]. A strategy is proposed by Samuel et al. which clusters and then indexes the tweets based on the emotions and emoticons present in the tweet [6].

Reviews also help us recognize market dynamics and tactics, as that may be achieved by nostalgic research as it encourages one to recognize popular items as it enables companies, enterprises, to use and grow accordingly. It can also be used by people themselves in general to search for which movie to watch, which laptop to purchase, but when we find spam reviews, a person does not know whether they are false or not in fact; however, they change our perspective. The author goes over this in a step-by-step style with various articles and outlines how a person can recognize the right feelings for other readers and distinguish between the true and the false reviews [7].

2.3 Sentiment Analysis Challenges and Techniques

Anees et al. mention the reviews of the product, the user got confused whether the reviews are right or wrong. If they are not sure about the reviews, they will take reviews from the people who have bought that product or they will assume whether the product is good or bad by seeing the rating of product [4]. The main objective is to find the suitable techniques to get the correct reviews. The information can be divided into subjective and objective forms. Objective means facts, and subjective means emotions which are used in sentiment analysis. Suggested machine leaning approaches, lexicon-based approaches, and hybrid approaches are been used [8]. The lexicon approaches divide the document into lexemes which are used to examine the sentences it is further divided into corpus-based and dictionary-based. Corpus finds

out the positive, negative, and neutral polarity of the sentences. Machine learning is divided into supervised and unsupervised learning. E-commerce dataset is used by Anees. Supervised learning requires the desired output with the actual output. In machine learning approach, they have used the Naïve Bayes. The related work is done in which the dataset is taken as 70% training data and 30% testing data.

$$P\left(\frac{c}{x}\right) = \frac{P(x/c)P(c)}{P(x)}$$
$$P(c|x) = P(x1|c)^* P\left(x2\middle|c\right) \ldots^* P(xn\middle|c)^* P(c).$$

The above equation is Naïve Bayes which is been commonly used to predict the accuracy of sentiments.

2.4 Sentiments Also Play a Major Role in Health Care

Abualigah Laith, et al. tell us about the sentiment in health care. There is some technique which can improve healthcare quality. The healthcare sentiments are related to the hospital performance, which hospital is best, and recovery process which can be taken from other patients. The technique tells that there is no need to ask other patients whether the hospital is good or bad. Past there were surveys of the different hospital, but they consume a lot of money and time. Abualigah Laith, et al. have taken a dataset in which the hospital list is there, and it has some option that reviews which is the best hospital for medication. Natural language processing involves a technique called emotion analysis mining; it recognizes the emotional meaning behind a paper picture. It is a common way for organizations to define and categorize an aggregate of feelings about a commodity, service or concept. Thus, the polarity of ambiguous words (context-dependent words) needs to be decided efficiently and effectively, and then, the aspect-based description produced. The author used k-nearest neighbour classifier in this paper to evaluate the polarity of the context-dependent terms [9].

Some resources are needed for this function such as polarized lexicon. Sentiment data mining in health care is not well studied, partially because patients are given some trust and an interpretation of their emotions, and often patients are using social media. These are inspired by the mining of product ratings in sentiment analysis. Next, they describe the root of the lexicon, using terms from the general realm of sentiment analysis and their polarity, and then they create a lexicon of medical emotion analysis based on a sample of drug feedback. Most views include terms of thought and have the same polarity in all circumstances. But there are several words of opinion called context-dependent terminology that in various situations have distinct polarities [10].

In this, the Arabic languages are been used the dataset of different categories in which different techniques are been implemented. The transmission of the natural language has many problems that may alter the nature of the expression of emotions

in many aspects. Some of the problems are linked to data form, while others are apparent to some sort of text analysis. In this many techniques are been applied to the Naïve Bayes, for example, is powerful and fast computing, without being affected by trivial features. It does presume individual characteristics, however [11].

2.5 Sentiment Analysis of Social Media, Politics, etc., Can Be Predicted by Deep Learning Techniques

The World Wide Web, such as social networks, groups, web pages, and blogs, creates vast volumes of data in the form of thoughts, feelings, viewpoints, and claims regarding different world activities, goods, brands, and policies. User emotions shared on the Internet have a great impact on readers, sellers of goods and policymakers. The unstructured type of social media data is expected to be processed and well-structured, and much focus has been paid to sentiment analysis for this reason [12]. Examination of emotion is referred to as an as text entity which is used to identify our emotions conveyed in various ways such as negative, good, favourable, and unfavourable. In the field of natural language processing (NLP), the problems for trend analysis are the lack of adequate labelled data. And the sentiment analysis and deep learning approaches have been combined to solve these problems, and deep learning models are successful because of their automatic learning ability.

As deep learning models can be a deep neural network, recursive neural network and many more other neural networks are been used. Via the use of deep learning models, this study presented adequate studies relevant to sentiment analysis. After reviewing all of these experiments, it is known that the interpretation of emotions can be done more effectively and reliably by using deep learning methods. As the study of emotions is used to forecast consumer attitudes, and deep learning models are more about modelling or mimicking the human mind, and they have more precision than shallow models. Deep learning networks are good than SVMs and normal neural networks because they have more hidden layers than normal neural networks with one or two hidden layers. The author introduces a system that focuses on clustering and indexing tweets, based on their geographical and temporal characteristics. The X-means clustering was used, which does not enable the user to enter the cluster number, but rather takes enter from the index of the tweets-created specified functions [13].

Many different forms of neural network are there, i.e. convolutional neural network, recursive neural network, deep neural network, recurrent neural network, deep belief network, hybrid neural network, and another neural network [14]. One is to capture people's opinions worldwide, which is called opinion mining or nostalgic research. It encourages them to consider the consumer need and help to produce the stuff people want and to isolate the items that are undesired or to solve a particular problem. This makes development at a much faster rate [15].

Deep learning networks perform automated extraction of features which does not require human interaction because it will save time and there is no need for

Table 1 Various approaches used for sentiment analysis

Authors	Approaches used	Dataset
Basiri et al. [1]	Lexicon approach	Hotel and movies dataset
Chen et al. [5]	Deep brief network and sentiment analysis	Yelp and Amazon dataset
Anees et al. [16]	Naïve Bayes	E-commerce dataset
Abualigah et al. [13]	NLP	Heath dataset
Zhang et al. [14]	CNN	Twitter dataset

software engineering. Sentiment analysis consists of various kinds of comments about problems. The ability to resolve differences in the process by making few modifications in the program itself requires a deep learning basic power feather. This approach often has some drawbacks relative to previous versions such as SVM. It needs massive datasets and is incredibly costly to train. These sophisticated models will train for weeks using computers fitted with costly GPUs [11] (Table 1).

3 Various Parameters

Lexicon approach: This approach uses a sentiment lexicon to explain the polarity of a textual material (positive, negative, and neutral). This methodology is more intuitive and can be applied quickly, as opposed to algorithms based on deep learning. Nonetheless, the downside is that it requires human intervention in the process of text analysis. The more popular the amount of material, the more notable the task would be for sifting through the noise, recognizing the meaning and separating valuable details from various sources of knowledge (Table 2).

Naïve Bayes: Easy term description based on 'theorem of Bayes.' This is a 'Bag of Words' technique for contextual interpretation of a substance (text described as set of its words, rejecting grammar, and word order while retaining multiplicity).

Table 2 Various parameters for sentimental analysis

Year	Approach use	References
2019	Lexicon	[1]
2020	Lexicon	[4]
2020	Naïve Bayes	[2]
2017	Deep learning	[16]
2019	DBNSA	[5]

Deep learning: Deep learning tailors a multilayer solution to the neural network's hidden layers. In conventional approaches to machine learning, features are described and extracted either manually or by using methods for selecting features. Nonetheless, features are automatically taught and extracted in deep learning models, gaining greater precision and efficiency.

Deep belief network and sentiment analysis (DBNSA): The DBN and sentiment analysis (DBNSA) approach is a discriminative classifier, which predicts the likelihood of rating from a WordNet sentiment analysis created word vector and then uses deep learning to train a model that predicts ranking.

4 Dataset

Three datasets are been used in which, preview dataset is manually labelled, and two other datasets of hotels and movies are been introduced [1].

The first collection of data is comprised of user comments on movies that are obtained from the Website of Naghdefarsi.com [17], and the second set of data includes comments on hotels reported on different Websites concerned. Aside from these latest datasets, they have named the current per view dataset [18] recently published in the Persian language for document-level SA. This data collection contains reviews of wireless appliances and will be compiled at Digikala.com in 2017 [16].

Next, this also contains three datasets, and these three show different results. Amazon dataset is used which is been collected by crawling the Amazon Website and camera category and its comment. 70% of the data for the deep learning model is for training and 30% for testing. Chen et al. use the Academic Challenge 5th round of Yelp data collection, which consists of over 1.5 million reviews, 36 600 users, and 61,000 companies [5].

There are 212,983 customer ratings for the hotel category in the Trip-Advisor data collection. The data collection is composed of 12,773 hotel travellers in tourism areas. The Trip-Advisor dataset is a Xml structure. The data collection of Trip-Advisor is obtained by browsing the Trip-Advisor Website and collecting only details from hotels and their reviews. The data collection for the Trip-Advisor is identical to that for the Amazon dataset [5].

The data is been collected from the e-commerce Website. Web scraper has been used to scrape comments from Amazon product URL and store them in spreadsheet form. The scraped comments are pre-processed to save time and energy for the computation [4].

The Arabic language suffers the lack of immense open databases for applications of AI and emotion analysis. This study launched a massive dataset, called BRAD, which is Arabic Dataset's biggest book reviews [19]. This dataset contains 490,587 inn surveys obtained from the Booking.com site record that includes the message of the survey in the Arabic language, the assessment by the commentator on a scale of 1–10 stars, and various attributes of the hosting/analyst. They make the complete

unequal dataset available just like a good subset. Six prevalent classifiers are used using Modern Standard Arabic (MSA) for evaluating the datasets [11].

ATwitter dataset containing 1269 images is chosen for experimental work, and back propagation is introduced. The photos are marked with Amazon Mechanical Turk (MTurk) and common crowd intelligence. Five employees were involved in creating sentiment mark for each graphic. On this dataset, the proposed model was evaluated and got better performance than existing systems. Results show that the device proposed achieves high efficiency without fine tuning on Flickr dataset [11].

5 Conclusion

As the analysis suggests about many different techniques and by which they get many different accuracies. An article, come across many different results, first is the decomposes of a long review into its constituent sentences and then detects the main target of each sentence. Finally, using the POS tags, the proposed method filters out all words except the potential terms, considering a comprehensive sentiment lexicon, and computes the polarity of the sentence. Moreover, five target identification strategies, including MOF, MGF, MSF, FOF, and LOF, are proposed to come up with the main target of the review. The author suggests a novel approach for estimating user ratings from user feedback by using a DBN and sentiment analysis. The sentiment analysis generates the function vector based on the good and bad opinions of products or services as stated by users in their comments. They also enforce noise reduction procedures that remove short posts, comments with no speech in them, and false rating comments. In this paper, author used a lexicon-based approach to measure feedback sentiment. Lexicon-based approach provides more specificity because they use a word. Analysis of emotions is a good way to help people get a decision and learn information. This approach attempts to examine the social Web, anywhere an identified issue will only reach the appropriate authority if they quickly find it. The best advice cannot be accessed via social media and different consumer materials. Sentiment analysis automates this process. Sentiment analysis is geared at gathering more knowledge to help consumers make the correct judgement about the analysed. To change this challenge, the techniques of sentiment analysis applied to data mining and machine learning. Author after reviewing all of these experiments, like convolutional neural network, deep belief neural network and many more, it is known that interpretation of emotions can be done more effectively and reliably way by using deep learning methods. As the study of emotion is used to forecast consumer attitudes, deep learning models are more about anticipating or imitating the human spirit, and deep learning models have greater precision than shallow models. Deep learning networks are different than SVMs and normal neural networks because they have more layers hidden than normal neural networks with one or two layers revealed. Deep learning networks can deliver training in both supervised and no supervised ways.

6 Limitations and Challenges

Chen et al. have not discussed social relationships and corresponding timeline comments, as user comments are influenced by the social relationships of the users. As consumers are influenced by their past interaction with related goods or services, timeframe is also a significant consideration [5]. Anees et al. [4] show some challenges that there can be multiple language input, fake input, emoticons, and sarcastic reviews. The transmission of natural language has many problems that in several ways will affect the presentation of the emotion analysis. Some of the problems are connected to data type, while others are obvious to any kind of text analysis. Compared to previous versions such as SVM, this system still has certain drawbacks. It needs massive datasets and is incredibly costly to train. These complex models will practice for weeks using fitted machines with costly GPUs [11].

References

1. Basiri ME et al (2019) Improving sentiment polarity detection through target identification. IEEE Trans Comput Soc Syst
2. Duwairi R, El-Orfali M (2014) A study of the effects of preprocessing strategies on sentiment analysis for Arabic text. J Inform Sci 40(4):501–513
3. Hajmohammadi MS, Ibrahim R (2013) A SVM-based method for sentiment analysis in Persian language. In: International conference on graphic and image processing (ICGIP 2012), vol 8768. International Society for Optics and Photonics
4. Anees AF et al (2020) Survey paper on sentiment analysis: techniques and challenges, No. 2389. EasyChair
5. Chen R-C (2019) User rating classification via deep belief network learning and sentiment analysis. IEEE Trans Comput Soc Syst 6(3):535–546
6. Samuel A, Sharma DK (2018) A novel framework for sentiment and emoticon-based clustering and indexing of tweets. J Inform Knowl Manage 17(02):1850013
7. Agarwal Y, Sharma DK, Katarya R (2019)Sentiment/opinion review analysis: detecting spams from the good ones! In: 2019 4th international conference on information systems and computer networks (ISCON), Mathura, India, 2019, pp 557–563. https://doi.org/10.1109/ISCON47742.2019.9036249
8. Digikala Internet Shop (2019) Accessed 23 Feb 2019. [Online]. Available: https://www.digikala.com/
9. Garg S, Sharma DK (2015)Feature based clustering considering context dependent words. In: 2015 1st international conference on next generation computing technologies (NGCT), Dehradun, 2015, pp 713–718. https://doi.org/10.1109/NGCT.2015.7375214
10. Garg S, Sharma DK (2016) Sentiment classification of context dependent words. In: Satapathy S, Joshi A, Modi N, Pathak N (eds) Proceedings of international conference on ICT for sustainable development. Advances in intelligent systems and computing, vol 408. Springer, Singapore. https://doi.org/10.1007/978-981-10-0129-1_73
11. Abualigah L et al (2020) Sentiment analysis in healthcare: a brief review. In: Recent advances in NLP: the case of Arabic language. Springer, Cham, pp 129–141
12. Maynard D, Funk A (2011) Automatic detection of political opinions in tweets. In: Extended semantic web conference. Springer, Berlin, Heidelberg
13. Samuel A, Sharma DK (2017) A spatial, temporal and sentiment based framework for indexing and clustering in twitter blogosphere, p361

14. Zhang Y et al (2016) Sentiment classification using comprehensive attention recurrent models. In: 2016 international joint conference on neural networks (IJCNN). IEEE, New York
15. Agarwal Y, Katarya R, Sharma DK (2019)Deep learning for opinion mining: a systematic survey. In: 2019 4th international conference on information systems and computer networks (ISCON), Mathura, India, 2019, pp 782–788. https://doi.org/10.1109/ISCON47742.2019.9036187
16. Nasr FM, Mohamed SE, Shaaban M, Hafez TAM (2017) Building sentiment analysis model using Graphlab. Int J Sci Eng Res 8:11551160
17. Naghdefarsi (2019) Accessed 23 Feb 2019. [Online]. Available: https://naghdefarsi.com/
18. Basiri ME, Kabiri A (2018) Words are important: improving sentiment analysis in the Persian language by lexicon refining. In: ACM transactions on Asian and low-resource language information processing (TALLIP), vol 17.4, p 118
19. Alayba AM et al (2018) Improving sentiment analysis in Arabic using word representation. In: 2018 IEEE 2nd international workshop on Arabic and derived script analysis and recognition (ASAR). IEEE, New York