A modern approach to sentiment analysis in big data: Methods, Tools, Applications, and open challenge

Abstract

In today's day and age, there are billions of volumes of textual content being generated everywhere. In-apps messages like WhatsApp, Telegram, social media sites like Facebook, Instagram, news publishing sites, google searches, and many other sources. All these sources are constantly generating huge volumes of text data every second. And because of these huge volumes of text data [NLP](https://www.sciencedirect.com/topics/engineering/natural-language-processing) becomes a vital resource in understanding the textual content. In this paper, the focus is on the popular NLP task of Sentiment analysis. Sentiment analysis is contextual mining of text which identifies and extracts subjective information in textual data. sentiment analysis proves to be an incredible asset for users to extract essential information and assists organizations with understanding the social sentiment of their brand, product, or service while monitoring online conversations. This paper investigates the different approaches and classification models used in the task of sentiment analysis.

* [**Previous**article in issue](https://www.sciencedirect.com/science/article/pii/S2666285X21000467)
* [**Next**article in the issue](https://www.sciencedirect.com/science/article/pii/S2666285X21000601)

Keywords

BERT

Deep learning

IMDB

LSTM

Machine learning

Naïve Bayes

NLP

Sentiment analysis

SVM

Transformers

Introduction

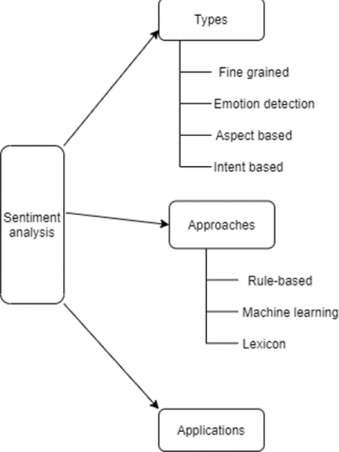
[Natural language processing](https://www.sciencedirect.com/topics/engineering/natural-language-processing) (NLP) is the ability of a computer program to understand human language as it is spoken and written also referred to as natural language. It is a component of artificial intelligence (AI). Sentiment analysis is a natural language processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis models focus not only polarity (positive, negative, neutral) but also on feelings and emotions (angry, happy, sad, etc.), urgency (urgent, not urgent) and even intentions (interested v. not interested).

By utilizing sentiment analysis, you measure how clients feel about various areas of your business without perusing a large number of client remarks at once. On the off chance that you have many inputs each month, it is inconceivable for one individual to peruse these reactions. By utilizing sentiment analysis and automating this interaction, you can undoubtedly bore down into various client fragments of your business and improve your understanding of sentiment in these segments.

While sentiment analysis is valuable, it's anything but a total swap for perusing survey responses. Frequently, there are important subtleties in the actual remarks in which sentiment analysis can assist you in further distinguishing which of these remarks you should peruse.

Taxonomy of research

This section briefly discusses the various aspects of sentiment analysis such as the types, approaches, [classification algorithms](https://www.sciencedirect.com/topics/engineering/classification-algorithm) and applications which are discussed in detail in the further sections of this paper. [Fig. 1](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "fig0001) depicts the taxonomy, which includes categories such as types, approaches, applications of sentiment analysis and their sub categories. This paper also explores the popular machine learning based approaches and the different classifiers used in them.



1. [Download : Download high-res image (153KB)](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr1_lrg.jpg)
2. [Download : Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr1.jpg)

Fig. 1. Taxonomy of Research.

Related works

In recent years there's been a significant rise in the sentiment analysis research or work using Machine learning, Deep learning and Transformer based algorithms. The works carried out using IMDB dataset have been taken into consideration to analyze their methodologies and algorithms and also compare the performances of each of them. Some of those works are:

Tirath Prasad Sahu [[3]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0003), utilized a methodology of determining the polarity using a lexical approach. Wherein the words comprising the document were considered and a few [classification algorithms](https://www.sciencedirect.com/topics/engineering/classification-algorithm) were applied to extract the collective polarity based on a sentiment score. [Machine learning algorithms](https://www.sciencedirect.com/topics/engineering/machine-learning-algorithm) such as Naïve Bayes, Decision Tree, KNN were applied on the IMDB dataset and the results were examined.

Arafat Habib Quraishi [[5]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0005), used machine learning and deep learning-based algorithms such as [SVM](https://www.sciencedirect.com/topics/engineering/support-vector-machine), [LSTM](https://www.sciencedirect.com/topics/engineering/long-short-term-memory), [GRU](https://www.sciencedirect.com/topics/engineering/recurrent) to perform sentiment analysis on the IMDB dataset. The performance metrics revealed that the deep learning or neural network-based approaches outperformed the classical machine learning based models in terms of binary classification

Saad Abdul Rauf [[6]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0006), implemented a transformer-based model such as BERT to analyze the polarity of the IMDB dataset using sentiment analysis. The BERT model performed exceptionally well on the given dataset as the result analysis revealed much better performance metrics than most of the previous machine learning and deep learning-based models.

Qizhe Xie [[8]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0008), proposed a new technique of Unsupervised Data Augmentation (UDA) which works very well while performing sentiment analysis tasks using BERT. This new technique implemented with the BERT model beats prior deep learning-based models by a clear margin.

T.Nikhil.Prakash [[15]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0016), [[16]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0017), [[17]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0018), focused mainly on three approaches, namely, machine learning, [lexicon](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/lexicon) based and hybrid approach and also three levels are discussed sentence level, document level, and topical level.

Zulfadzli Drus [[19]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0019), this research worked showed that most of the articles considered opinion-lexicon method to analyze text sentiment in social media.

Types of sentiment analysis

To understand the sentiments of people, there are different types of sentiment analysis used in the market. Apart from normal opinions – positive, negative, or neutral, other types of sentiment analysis help in understanding people's inner feelings, their actual intentions, and emotions.

Fine-grained sentiment

This is one of the most simple and common ways of understanding the customers' sentiments. This analysis gives us an understanding of the feedback received from customers.

While analyzing the sentiments, readily available categories like positive, neutral and negative are used. Providing a rating option from 1 to 5 is another way to scale the feedback given by your customers. Most e-commerce sites use this technique to know the sentiments of their customers.

Emotion detection sentiment analysis

This is a more refined method of detecting the feeling in a piece of text. This kind of analysis aids to detect and understand the emotions of the people. Emotions like anger, sadness, happiness, frustration, fear, panic, worry may all be included.

The upside of utilizing this is that an organization can also understand why a customer feels a specific way, but understanding the sentiments of people using emotion detection is very difficult as people use a collection of words having a different sense of meaning such as sarcasm.

Aspect-based analysis

Sentiment analysis of this type is more focused on the aspects of a particular product or service.

Aspect based sentiment analysis is essential as it can support organizations in automatically sorting and analyzing customer data, automating the processes like customer support tasks allows us to gain significant insights on the fly.

Aspect-based sentiment analysis empowers organizations to zero in on the parts of their products or administrations that their clients are griping about and helps them in fixing those issues progressively. Complaints such as glitches or major bugs in some new software applications can also be addressed.

Intent-based sentiment analysis

Intent classification refers to the automatic classification of textual data which is based on the customer's aim. An intent classifier can naturally dissect the writings and reports and classifies them into intents like Purchase, Downgrade, unsubscribe etc. This proves helpful to comprehend the intentions behind a large number of the client's questions, automates measures, and acquires significant experiences.

Intent classification empowers organizations to be more customer-friendly, especially when dealing with areas such as customer support and sales. It helps them in reacting to leads faster and handling large volumes of inquiries.

Sentiment analysis approaches

Rule-based approach

The main strategy is rules-based and utilizes a dictionary of words named by sentiment to decide the sentiment of a sentence. Sentiment scores regularly should be joined with extra principles to relieve sentences containing negations, sarcasm, or dependent clauses [15-[18](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0018)].

The [NLP](https://www.sciencedirect.com/topics/engineering/natural-language-processing) techniques which are included in the rules are:

•

Stemming, tokenization, part-of-speech tagging

•

[Lexicons](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/lexicon).

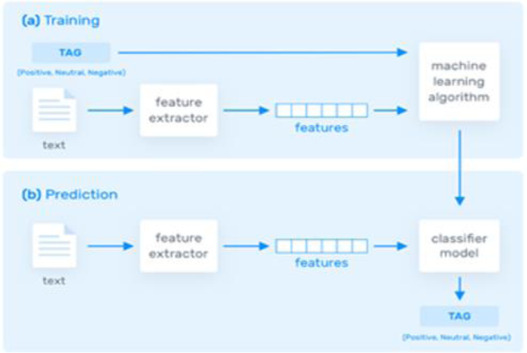
Systems involving rule-based approaches are quite simple since the sequential merging of words is not considered. superior processing methods can be utilized and the latest rules can be affixed to support newer forms of expression and vocabularies. But, the addition of new rules can influence previously obtained results and can cause the entire system to become extremely complicated. Frequent fine-tuning and maintenance are required by rule-based systems hence, will also require financing at frequent intervals.

Machine learning approach

[Machine learning techniques](https://www.sciencedirect.com/topics/engineering/machine-learning-technique), don't depend on manually crafted rules, but on machine learning procedures. A sentiment analysis task is normally modelled as a classification problem, where the classifier a text data and it returns a class such as positive, negative or neutral [[10]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0010).

•

In the training process [Fig. 2](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "fig0002)(a), our model learns to compare a particular input data to the respective output data on the basis of test examples utilized for the training process. The feature extractor transforms the textual input into a features vector. Feature tag and vector pairs are then supplied into the algorithm to produce a model.



1. [Download : Download high-res image (266KB)](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr2_lrg.jpg)
2. [Download : Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr2.jpg)

Fig. 2. Machine Learning Approach.

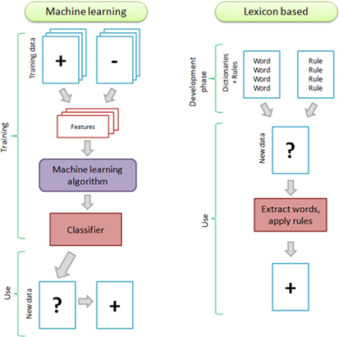
•

In the prediction process depicted in [Fig. 2](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#fig0002)(b), the feature extractor transforms hidden textual inputs into feature vectors. These vectors are then supplied to the model, generating prediction tags for the corresponding vectors.

Lexicon based approach

This method figures the sentiment orientation of the entire text or set of sentences(s) from semantic direction of lexicons. Semantic direction can be positive, negative, or neutral. The word reference of lexicons can be made manually as well as automatically generated [[11]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0011).

[Fig. 3](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "fig0003).



1. [Download : Download high-res image (225KB)](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr3_lrg.jpg)
2. [Download : Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr3.jpg)

Fig. 3. Machine Learning Approach vs [Lexicon](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/lexicon) Based Approach.

The methodology for the sentiment categorization task was executed as follows. Firstly, all training weights of text data and the categorized text is calculated. The entire textual data is then stored in a 1-D emotion field. Then the mean weights of the training text data per sentiment category were identified. The categorized text belonged to the category found closer in the 1-D emotion field [[19]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0019).

The above image (3) describes the process of machine learning approach and lexicon-based approach.

Advantages and Limitations of All the Approaches

In the below [Table 1](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "tbl0001) advantages and limitations between Rule-Based approach, Machine Learning approach and Lexicon-Based approach have been discussed.

Table 1. Advantages and Limitations of approaches

|  |  |  |
| --- | --- | --- |
| **APROACHES** | ADVANTAGES | LIMITATIONS |
| **Rule Based Approach** | Training data not required High precision Can be a good way to collect data as one can start the system with rules and let data come by naturally as people use the system. | Lower recallDifficult and tedious to list all the rules |
| **[[1]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0001) Machine Learning Approach** | Dictionary is not required. Exhibit the high precision of classification. | Classifier trained on the textual data in a single field much of the time doesn't work with different fields. |
| [**[1]**](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0001)**Lexicon Based Approach** | Named information and the method of learning isn't needed. | Requires amazing semantic assets which isn't generally accessible. |

Table 2. Accuracy of all the Discussed Models.

|  |  |  |
| --- | --- | --- |
| **Type of classification model** | Classification model | Accuracy |
| Empty Cell | [[3]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0003) Naive Bayes | 54.77% |
| **Conventional Machine learning** | [[3]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0003) Decision Tree | 87.53% |
| Empty Cell | [[5]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0005) Support Vector Machine (SVM) | 88.3% |
| Empty Cell | [[3]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0003) K-Nearest Neighbor (KNN) | 88.86% |
| **Deep Learning based** | [[5]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0005) LSTM | 88.5% |
| Empty Cell | [[5]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0005) GRU | 89.0% |
| **Transformer based** | [[6]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0006) BERT | 89.5% |
| Empty Cell | [[8]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0008) BERT Large with UDA | 95.22% |

Machine Learning Classification Algorithms

To compare and analyze the performance of each model, the classifiers are selected from each category such as conventional classifiers (7.1), deep learning-based classifiers (7.2) and transformer-based classifiers (7.3).

Conventional Machine learning classifiers

The machine learning models which are commonly used for text classification, which are quite simplistic in their methodology and sort of take a linear approach in classifying the data are considered first for the analysis. The classifiers such as Naïve Bayes, KNN, [SVM](https://www.sciencedirect.com/topics/engineering/support-vector-machine) and Decision Tree are considered.

Naive Bayes (NB)

The Naive Bayes algorithm is one of the intuitive methods among classification algorithms. It is a simple algorithm that makes use of the probability of every feature per category to get respective predictions. This algorithm runs great for the categorization of textual data. It is based on the [Bayes Theorem](https://www.sciencedirect.com/topics/engineering/bayes-theorem) which is used to describe the probability of an event based on its prior knowledge [[9]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0009).

The Naive Bayes classification model presents a simplistic method with a clear explanation to describe, utilize and learn probability knowledge. Naive Bayes is also a part of supervised learning techniques [[3]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0003).

Bayes' theorem is represented mathematically with the following [equation (1)](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "eqn0001):(1)P(B)=P(A)\*P(A)P(B)where A and B are events and P(B) ≠ 0.

(A|B), a [conditional probability](https://www.sciencedirect.com/topics/engineering/conditional-probability), is the probability of observing an event A given that B is true.

P(A) and *P*(B) are the probabilities of witnessing A and B without considering one another.

(B|A): the probability of witnessing event B provided A is true

K-Nearest Neighbor (KNN)

The k-nearest neighbor (KNN) algorithm is a simple, supervised [machine learning algorithm](https://www.sciencedirect.com/topics/engineering/machine-learning-algorithm) that can be used to solve both classification and regression problems [[4]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0004).

KNN is an algorithm that classifies new objects based on the k-value. The k-value should be less than the volume of training data. The target that has to be categorized can be decided by using the training data distance nearest to the target [[2]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0002).

This task is done using the [Euclidean Distance](https://www.sciencedirect.com/topics/engineering/euclidean-distance) formula, shown below in [equation (2)](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "eqn0002):(2)D(a,b)=∑k=1d(ak−bk)2

Where D – represent the scalar distance between the respective vectors, a –represents test data, b – represents train data.

After calculating the Euclidean Distance, the result is sorted and selected as per K value in order of data quantity to determine the classes.

Support Vector Machine (SVM)

Supervised machine learning algorithms such as SVM examines data and identifies patterns utilized for categorization. SVM possesses the ability to identify separate [hyperplanes](https://www.sciencedirect.com/topics/engineering/hyperplanes) which helps in maximizing the margin between the various classes [[13]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0013).

The methodology for classifying using the SVM algorithm starts by transforming text data into weights. These weights are frequently fused to form TF-IDF values, by just multiplying them collectively [[2]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0002).

Positive or negative sentences can be determined by calculating the hyperplane using the [equation (3)](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "eqn0003) and [equation (4)](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "eqn0004) given below:(3)f(∅(x))=sign(w.∅(x)+b)Where f(∅(x)) – represents the result of testing data categories, w –represents weights, b – represents bias, Φ(x)–represents testing data of kernel calculations(4)K(x,xi)=(x.xi+1)2Where K(x,xi) – kernel, x – train data, xi – train data to i.

By determining the hyperplane, calculation of the test data is performed based on the weights of the test data by considering positive or negative classes [[2]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0002).

Decision Tree (DT)

Another supervised learning algorithm such as Decision Tree is used when non-linear datasets need to be handled efficiently. The decision tree technique is pretty effective in building classification models using the available data. Decision tree varieties are popularly used with logical techniques [[14]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0014).

The Decision Tree is a flowchart like arrangement that matches a tree. The class label is represented by each [leaf node](https://www.sciencedirect.com/topics/engineering/leaf-node) or [terminal node](https://www.sciencedirect.com/topics/engineering/terminal-node) and the [root node](https://www.sciencedirect.com/topics/engineering/root-node) is represented by the node present at the peak of the decision tree [[14]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0014).

The equation for the Entropy of decision tree is represented in [equation (5)](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "eqn0005):(5)E(D)=∑i=1n−pc(i)(pc(i))Where, pc(i) is the probability of class C(i) in a node.

E(D) or also called entropy of D is the measure of disorder of the considered samples.

Deep Learning Based Classifiers

The Deep learning models follow the [neural network approach](https://www.sciencedirect.com/topics/engineering/neural-network-approach) which is inspired by the biological neural network observed in human beings. These models learn from the provided data similar to the human brain by adjusting a ‘weight’ parameter and training them continuously over several epochs [[13]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0013). The deep learning models such as [LSTM](https://www.sciencedirect.com/topics/engineering/long-short-term-memory), [GRU](https://www.sciencedirect.com/topics/engineering/recurrent) which are known to perform well with sequence and textual data are considered for the analysis.

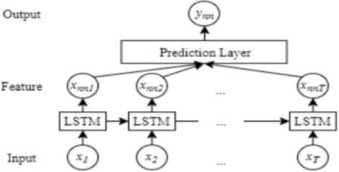
Long Short-Term Memory Model (LSTM)

Long Short-Term Memory Model (LSTM) units are units of [Recurrent Neural Network](https://www.sciencedirect.com/topics/engineering/recurrent-neural-network) (RNN). LSTM networks are used for classification or prediction based on time series data. It can deal with exploding and vanishing gradient problems. One of the major problems of RNN is long-term dependency. LSTM can avoid such long-term dependencies [[6]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0006).

A regular LSTM unit consists of a cell, input gate, output gate and forget gate. The input gate determines the amount of the previous sample which is stored in memory. The output gate controls the amount of data sent to the next layer, the forget gate regulates the rate of loss of stored memory and determines what data can be dropped from the state of the cell.

LSTM's [weight changes](https://www.sciencedirect.com/topics/medicine-and-dentistry/body-weight-change) through the input gate, the forget gate and the output gate, thereby avoiding the issue of gradient vanishing or exploding gradient.

The [figure 4](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "fig0004) represents the LSTM Network [[6]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0006).



1. [Download : Download high-res image (92KB)](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr4_lrg.jpg)
2. [Download : Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr4.jpg)

Fig. 4. [LSTM](https://www.sciencedirect.com/topics/engineering/long-short-term-memory) Network.

Gated Recurrent Unit (GRU)

Gated Recurrent Unit (GRU) is a simplified version of Long Short-Term Memory (LSTM) model [[6]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0006). Even though GRU has fewer parameters, the model is able to efficiently capture long term dependencies between sequences. Therefore, GRU is comparable to LSTM in terms of performance and computational efficiency[[12]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0012). Gated Recurrent Unit (GRU) calculates two gates called update and reset gates which control the flow of information through each hidden unit. The [Figure 5](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "fig0005) shows the operation of LSTM and GRU cells [[6]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0006).

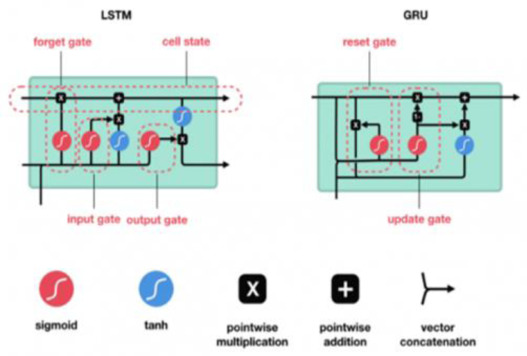
Transformer based models

Transformer models are a kind of deep learning models that follow the attention mechanism and sequence to sequence architecture. Attention mechanism works by looking at input sequences and by deciding at each step which remaining parts of the sequences are important [[18]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0019), [[19]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0020), [[20]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0021). One such model known as BERT which is specifically fine-tuned for [NLP](https://www.sciencedirect.com/topics/engineering/natural-language-processing) tasks is discussed.

BERT

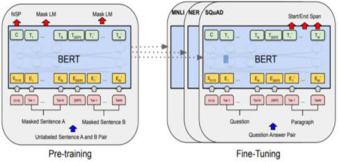
BERT (Bidirectional Encoder Representations from Transformers) is a Natural Language Processing Model proposed by researchers at Google Research in 2018[7]. BERT uses many previous NLP algorithms and architectures and architectures such as semi-supervised training, OpenAI transformers, ELMo Embeddings, ULMFit, Transformers. BERT is an Encoder stack of transformer architecture. A transformer architecture is an encoder-decoder network that uses self-attention on the [encoder side](https://www.sciencedirect.com/topics/engineering/encoder-side) and attention on the [decoder](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/decoder) side.

BERT consists of two steps [figure 6](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "fig0006) [[7]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0007): pre-training and fine-tuning. During pre-training, unlabelled data is used to train the model over various pre-training tasks. For fine-tuning, firstly pre-trained parameters are used to initialize BERT, and all of the parameters are fine-tuned using labelled data from the downstream tasks. Each downstream task has separate fine-tuned models, despite the fact that they are initialized with similar pre-trained parameters.



1. [Download : Download high-res image (323KB)](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr5_lrg.jpg)
2. [Download : Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr5.jpg)

Fig. 5. [LSTM](https://www.sciencedirect.com/topics/engineering/long-short-term-memory) vs [GRU](https://www.sciencedirect.com/topics/engineering/recurrent) cells.



1. [Download : Download high-res image (194KB)](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr6_lrg.jpg)
2. [Download : Download full-size image](https://ars.els-cdn.com/content/image/1-s2.0-S2666285X21000327-gr6.jpg)

Fig. 6. Two Step Process in BERT.

Comparison of The Classification Models

The [classification algorithms](https://www.sciencedirect.com/topics/engineering/classification-algorithm) discussed above have been applied to the IMDB movie review dataset to analyze the performance of the algorithms for the task of sentiment analysis.

The accuracy metric obtained based on the research work of Tirath Prasad Sahu[3], Arafat Habib Quraishi[5], Qizhe Xie[8], Saad Abdul Rauf[6] on the IMDB dataset is shown in [Table 2](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "tbl0002) .

Applications Of Sentiment Analysis

Brand Sentiment Analysis Management Of Reputation and Social Media Monitoring

Reputation Management and Monitoring of Brand are the basic utilization of sentiment analysis over different organizations. Nothing unexpected - knowing on what premise clients notice your organization/stock/administration are for the most part similarly valuable for innovation organizations, marketing agencies, design brands, news, and paper organizations, a few different organizations.

As a general rule, utilization of sentiment analysis brings extra adaptability and the inward delivery of the organization and its items. It empowers organizations to:

•

Track the comprehension of the organization by clients.

•

Pointing out the specific details of the organization towards the organization.

•

Company should discover the current example and the patterns.

•

Monitor what influencers are recommending about the brand.

Model: KFC

KFC is an extraordinary model on the best way to apply sentiment analysis for an organization becoming well known is Kentucky Fried Chicken. Before, Kentucky Fried Chicken was stuck while the rival(s) were progressing and building up itself with the impression of energizing food and bright encounters.

In this way, instead of attempting to substantiate themselves in the jam-packed corner, Kentucky Fried Chicken had chosen to utilize the general impact of the brand. Kentucky Fried Chicken began utilizing the prevalence of images, films, and music to expand the organization's position.

This strategy makes unsurprising contact about the brand that is extended by the mainstream society support. Thus, clients show interest in the organization and the end will be headed to utilize the item at last.

Market Research, Competitor Analysis

Besides organization comprehension and client assessment research, the research market is likely the most common well-known field of sentiment analysis uses.

Comprehending this sentiment analysis isn't the primary medium for market research. Regardless, it's anything but an additional point of view about the market and gives a few accommodating experiences about how items are seen from the client's perspective.

Other than that, we can utilize a related path to analyze the rival(s) and their retailing endeavors.

These are the things that should be possible utilizing sentiment analysis:

•

View important data across different stages

•

Customer-produced content

•

News reports

•

Expert content

•

Rival(s) content

•

Obtain different bits of knowledge on different scales

•

Common understanding

•

Knowledge of the specific highlights

•

Produce results in real-time

•

All the intel of the business ought to be kept

Sentiment analysis can be created on the prerequisites and wants of the clients and help to improve your value so it would break the very best outcomes.

Model: Apple

How Apple shows its merchandise and gets them on the business stage is the best model for sentiment analysis applications for the benefit of competitor analysis and research market.

Think about how conveniently solid marks of the item's fit the overall troubles of different pieces of the client.

Like:

•

Poor plan - you don't need to stress while utilizing our items.

•

security - we don't screen every one of your developments.

•

battery life - we have a decent battery to run the entire day.

Product Analytics

The utilization of sentiment analysis in the analysis of products originates from reputation management. Thoughtfully, it is the same as brand observation. However, the brand refers to, explicit remarks and comments concerning Employments of sentiment analysis in the analysis of products [[18]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0019), [[19]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0020), [[20]](https://www.sciencedirect.com/science/article/pii/S2666285X21000327#bib0021).

These sorts of experiences are vital during essential stages with Memetic Voter Patterns when you need these products to be tried by genuine clients and make them as polished as could be expected.

During this progression, the most popular strategy to utilize sentiment analysis is to collect and characterize reactions for betterments [[21](https://www.sciencedirect.com/science/article/pii/S2666285X21000327" \l "bib0022)] .

Sentiment analysis calculations can show what sort of input is received from clients and when was this criticism received.

Model: Google

The best display of how sentiment analysis uses offers product improvement can be seen in Google's yield. We should think about Chrome for instance.

Chrome's improvement team is persistently noticing user's audits, nevertheless, it is straightforward or not.

They are not perusing client criticism as a criticism from the client but instead as a number of its parts:

•

sentiment of the client

•

Remarks of the definite highlights of the product – nonetheless it is adaptability, expansions, security, or UI.

•

Emotional need, and ideas concerning the product in general and its precise subtleties.

Voice of the Customer Analysis

Precise objective crowd division and ensuing incentive definition are among the vital components of powerful business activity. You need to realize where are you focusing on with what.

On the other side of the range, you need to keep the hand on the beat of your customer to stay applicable and keep your item popular.

In the actual focus of the two exercises is a comprehension of the "Voice of the customer".

In any case, one doesn't just catch and study the voice of the customer. It is spread around the various stages and introduced in an assortment of clashing structures. This should be figured out quite clear.

Customer Sentiment Analysis calculations are equipped for catching and examining the voice of the customer with a lot greater precision.

Model: TripAdvisor

A genuine illustration of VOC analysis done right is TripAdvisor.

The association applies point of view-based sentiment analysis to make the most out of the enormous proportion of data it makes. The perspective-based procedure grants to isolates the practical focus concentrations with respect to customer input and the genuine assistance.

As the result, sentiment analysis gives an additional perspective on various bits of the business movement, which grants us to fathom what the planned vested party needs, figures, feels can be improved, and so forth

Eventually, this adds to the further clean of the help and strengthening of customer commitment by furnishing them with what they need.

Customer Support - feedback analysis

Customer Feedback Support is one of the major components of sentiment analysis utilization in real-time.

Sentiment analysis can be performed in various ways:

Customer's review of the product:

1

The common knowledge of the product - whether it is clear or not;

2

Aspect-based - concerning selective components of the product;

3

Reaction to the Service - whether it is useful or not. May also involve more complicated analysis concerning critical features such as response time or quality of communication;

Intent Analysis for automation process - so that regular doubts will be managed automatically, such as questions that are asked frequently and product information.

Customer management and Workflow management. For example, we have a dissatisfied user - his ticket is prioritized to be finished at the earliest.

Conclusion

This paper reviewed the various applications, approaches and classification models used for the task of sentiment analysis. We have discussed the advantages and disadvantages of the different approaches such as Rule-based, Machine learning approaches used for sentiment analysis as well as compared the performances of the classification models for the same. The accuracy results for the models based on the IMDB dataset illustrated that machine learning approaches such as

[SVM](https://www.sciencedirect.com/topics/engineering/support-vector-machine), [GRU](https://www.sciencedirect.com/topics/engineering/recurrent) and BERT showed exceptional accuracy. Notably, the more recent models such as GRU and BERT showed accuracies exceeding conventional classification models such as Naive Bayes, Decision Tree etc.

We have also discussed various applications as well as use cases of sentiment analysis.

In the modern era, as the volume of data being generated keeps increasing day by day, more powerful classification models such as BERT will have to be utilized to help the stakeholders such as Businesses and Social media platforms to improve the understanding they have of their users and help them in their decision-making process.

ABSTRACT The development of IoT technologies and the massive admiration and acceptance of social media tools and applications, new doors of opportunity have been opened for using data analytics in gaining meaningful insights from unstructured information. The application of opinion mining and sentiment analysis (OMSA) in the era of big data have been used a useful way in categorizing the opinion into different sentiment and in general evaluating the mood of the public. Moreover, different techniques of OMSA have been developed over the years in different data sets and applied to various experimental settings. In this regard, this paper presents a comprehensive systematic literature review, aims to discuss both technical aspect of OMSA (techniques and types) and non-technical aspect in the form of application areas are discussed. Furthermore, this paper also highlighted both technical aspects of OMSA in the form of challenges in the development of its technique and non-technical challenges mainly based on its application. These challenges are presented as a future direction for research. INDEX TERMS Opinion mining, sentiment analysis, big data, applications, opinionated data, social media, online social network. I. INTRODUCTION The 21st century has witnessed a torrential flow of data. The data has sprung massively in various fields over the last two decades, which has led to the birth of big data [1]. Moreover, the influx of technology in the digital world has opened the doors for the development of big data. Citizens of the world are now becoming technology savvy with devices ranges from digital sensors, communication tools including social media applications, and actuators and data processors [2]. For instance, organizations capture the mushrooming volume of transactional data, through which trillions of bytes of information are generated regarding aspects from suppliers to customers. The physical world has millions of network sensors embedded in devices like smart phones, smart energy meters, automobiles, and industrial machines [3]. Such advances in digital sensors and communication technologies have led to the development of the Internet of Things (IoT) [4]. With such a development, social networking sites and communication devices like smart phones, laptops, and PCs allow individuals to interact with one another to create massive amounts of big data [3]. For instance, Twitter’s wide network of 467 million users generates 175 million tweets on a daily basis [5]. Similarly, the amount of space needed to store one second of a high definition video is 2000 times more than the space needed to store a page of plain text [3]. Furthermore, according to the International Data Corporation report in 2011, the world is already generated about 1 zettabyte (ZB) of data, and the rate at which this amount is growing has been exploding; the amount of data grew to 7ZB by the end of 2014 [6]. Moreover, by 2020, the amount of data generated is expected to reach 44ZB, with at least half of them being textual data [7] that is generated through social media technologies like Facebook, Twitter, and mobile instant messaging apps such as WhatsApp and Telegram. It has been determined that 500 million tweets are sent each day, while 40 million of those are shared daily. Meanwhile, it is estimated that 4.3 billion messages on Facebook are posted with 5.75 billion likes on a daily basis. Moreover, it is expected that the amount of data will continuously grow because of the influx of digital technologies that have already sprung up in the digital era [1]. VOLUME 6, 2018 2169-3536 2018 IEEE. Translations and content mining are permitted for academic research only. Personal use is also permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications\_standards/publications/rights/index.html for more information. 37807 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges The extensive use of technologies and astounding flow of data over the years has also aided in the escalation of big data business analytics. As the term suggests, it refers to two components: big data and business analytics. According to the McKinsey Global Institute, ‘‘Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze’’ [3]. However, big data is generally defined through the key characteristics of volume, variety, and velocity [8]. 1. Volume represents the quantity of data that utilizes massive storage space or entails sizeable and considerable number of records [9]. For instance, Wal-Mart manages to store 2.5 petabytes of information, Tesco generated 1.5 billion new items of data on a monthly basis, and Dell developed a database which can handle 1.5 million sales and advertisement records [3], [10]. 2. Variety refers to the data generated from a range of sources and in varying formats [9]. The sources can be in the form of sensors, social media sites, web technologies, mobile phones, etc. The data format can include web logs, unstructured data like audio, videos, images and sensory data from RFID devices, or other smart sensors. 3. Velocity indicates the frequency with which data is produced from different sources [9]. The data can be generated occasionally, frequently, and/or on a realtime basis. On the other hand, analytics refers to the capability of the firm in applying statistics, econometrics, mathematical, simulation, and optimization tools [11] to obtain insights from data and employ data-driven decisions in the organizations [12]. Hence, companies in the majority of sectors are now focusing on acquiring profound information from the massive quantity of already-available data to gain competitive advantage [13]. Big data offers great challenges to organizations because of the nature of its complexity. At a fundamental level, organizations face the challenge of handling and storing a gigantic amount of data [14]. Moreover, there is a need to improve method for dealing confounding amount of raw data in a variety of forms [15]. Additionally, it is essential to design scalable data storage in order to acquire and retrieve meaningful information efficiently. High speed networking infrastructure can also consume less power when processing the data [2]. Furthermore, distributing information across many systems is another challenge which is critical in processing huge amounts of data from different datasets in a reasonable period of time. Cloud computing programming is one way to overcome this issue [16]. Apart from all these issues, privacy, security, accessibility, and deployment are additional issues that decision makers must consider before exploiting the vast benefits of big data to gain competitive advantage [14]. We commonly think about what other people think in our decision-making process [1]. Prior to the advent of the Internet, many of us relied on friends and families for product or service recommendations, voting views during local elections, or information when buying a product. The Internet eases our efforts to get opinions and experiences of those that are neither on our personal contact list nor in our professional networks. The amount of opinions and comments on the Internet has not only risen during the last decade or so, but they are also readily available to the strangers. However, opinion mining and sentiment analysis face challenges such as making query classification, classifying documents that contain reviews or explicitly opinionated material, determining sentiments based on the content, and finally presenting the sentiment information in a reasonable summary [1]. The main aims of this paper are as follows: ➢ To provide a comprehensive systematic literature review and discuss both the technical aspects of OMSA (techniques, types) and non-technical aspects in the form of application areas. ➢ To highlight both technical (related to the development of sentiment technique) and non- technical challenges (related to the application of Sentiment analysis). These challenges are presented as future research direction. A. AN OVERVIEW OF TEXT MINING AND ANALYTICS Since an enormous amount of data has emerged over the years at a staggering rate, there is a need to incorporate some sort of analytics to gain meaningful insights from the raw and unstructured data in the form of text, images, and videos. Text mining is one of the approaches that are the predecessors to text analytics. Text mining uses natural language processing, knowledge management, data mining, and machine learning techniques to process text documents [17]. Text analytics, while similar to text mining in terms of method, usually deal with a bigger amount of data to extract and generate useful non-trivial information and knowledge [18]. Text mining/analytics are originally conducted for two purposes. The first purpose is to analyze people’s sentiment on an issue or phenomenon. Hence, sentiment analysis goes through a huge amount of textual data to identify people’s attitudes, thoughts, judgments, and emotions on an issue [19], [20]. The second purpose is to assess people’s opinion on a product, person, event, organization, or topic from a user or group of user perspectives. Similar to sentiment analysis, opinion mining is a natural language processing task that employs an algorithmic technique to recognize opinionated content and categorize it into positive, negative, or neutral polarity [21]. Nonetheless, the application of text mining/analytics has been extended to other areas of human computer applications, and the applications are growing with the growth in big data analytics. B. AN OVERVIEW OF OPINION MINING AND SENTIMENT ANALYSIS An opinion refers to a person’s or group’s sentiment or views, emotions, and attitudes about a product, service, occasion, or other topic present in the environment. Like sentiment analysis, opinion mining is also grounded on the algorithmic 37808 VOLUME 6, 2018 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges technique [21]. Covering a huge variety of pubic opinions, [22] have argued that opinion can be classified into three main types: regular opinions, which refer to a single entity only; comparative opinions, which compare or contrast more than one entity; and suggestive opinions, which suggest a single or multiple entities. The regular opinion is mainly used to identify a positive or negative outlook of a particular product [23]. On the other hand, comparative opinions help in elucidating the association among multiple entities and are mainly used for competitive intelligence [23]. However, there is a dearth of literature concerning the identification of comparative sentences that is being used for the comparison of multiple entities. Recently, suggestive review has been introduced in the field of opinion mining [24]. The extraction of these types of opinions from text can be utilized for various application areas in the field of business, engineering, medical science, and e-learning. It can be offered to various online communities for their assistance as well [22]. Similarly, private statements of individuals are called sentiments, which comprise thoughts, opinions, attitudes, views, judgments, and feelings. These are commonly gathered by conventional scientific methods [19], [20]. [25] pronounced the feelings that are expressed in language by using subjective expression. The sentiments can be analyzed through the machine learning technique, which can be further classified into supervised and unsupervised, using a lexiconbased approach, using a keyword, and using a concept-based technique [26]. Recently, research on sentiment analysis has focused on multiple modalities such as in speech and video as opposed to earlier work that focused on unimodality related to text [27], [28]. Sentiment analysis tackles many NLP subtasks, including aspect extraction [29], subjectivity detection [30], named entity recognition, and sarcasm detection [31]. In most cases, the main objective of sentiment analysis is to unearth people’s opinions to gain meaningful insight about products or services. Its aim is to exhibit useful information to both customers and manufacturers. It is established that both manufacturer and customers look upon summarized opinions instead of detailed reviews. Hence the opinions that are categorized on positive, negative, or neutral sentiments are useful for both parties in making the right call [32]. Despite the large number of studies on opinion mining and sentiment analysis techniques, the impact they have on people has been less explored. There has been great emphasis on the techniques used and less on how people can benefit from the findings. Hence, this study aims to investigate the human element in opinion mining and sentiment analysis research. To achieve this aim, we will systematically review the relevant literatures that have employed both approaches. The study offers several contributions. The first and foremost significance of this study is to refocus the study of opinion mining and sentiment analysis to both technical and non-technical challenges. Secondly, it places emphasis on the areas of opportunity by looking at the trends of application coverage that would offer some potential areas for research. Thirdly, the paper presents information on different datasets that were used in opinion mining and sentiment analysis studies that future researchers could use in their research. The remaining section of the paper is organized as follows. First, the study talks about the method employed to achieve the research objectives. Then we present the findings of the study. Next, the paper highlights the commonly used dataset in literature. Later sections discuss different review methods for sentiment analysis and highlight the application area for sentiment analysis. Finally, the study emphasizes the current open challenges of big data sentiment analysis. II. METHODOLOGY The main goal of this paper is to develop a deep understanding of the various opinion mining and sentiment analysis approaches performed on human applications of text analytics. This study advocates the ways applications are present and utilized in many areas in the society. The technique used in the study is the systematic literature review. A systematic review is completely based on an evidently framed question, presents relevant studies, evaluates their findings, and summarizes the evidence by means of clear methodology. This unambiguous and methodical approach makes systematic reviews different from the traditional reviews. III. REVIEW OF OPINION MINING AND SENTIMENT ANALYSIS Literature reviews are rooted in medical science, which has been categorized as a critical approach mainly applied where evidence is important [2]. They involve a stringent process of finding, selecting, and examining secondary data [33], [34]. The synthesis of evidence from the current literature can create new knowledge in the current studies, which is as significant as conducting new research [35]. Rousseau et al. [36] maintained an argument that systematic reviews are different from tradition reviews in that systematic literature reviews are comprehensive in nature, use transparent and unbiased analysis, and apply certain criteria for interpretation of the findings that provided in the previous literature. In addition, systematic literature reviews mainly focus on objectivity and reproducibility of results [34]. The process of review starts with framing the questions and conducting a systematic and step-by-step process and applying a replicable method to answer these questions [34]. Thus, the evidence generated from the rigorous approach of identifying, selecting, and analyzing the data can have a significant impact on the body of knowledge collected, but the supreme concern of this practice is synthesizing the results produced through this systematic process [33], [34]. The methodology used in the study is a five-step process shown in Figure 1, as proposed by [33]. It is systematic in nature, clear and reproducible, and involves identifying, examining, synthesizing, deducing, and reporting the evidence from the existing sentiment analysis and opinion mining literature. VOLUME 6, 2018 37809 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges FIGURE 1. Research methodology of systematic literature review. A. QUESTION FORMULATION A deep and insightful literature review should start with the development of a clear understanding about your objectives [35]. Therefore, to ascertain this, we clearly formulated and considered research questions to evade doubts in our study [36]. The purpose of the paper is to discuss the methodological and application side of opinion mining and sentiment analysis and explore whether the intervention of opinion mining and sentiment analysis would be applicable to humans or in an organization as a whole. Hence the purpose of our systematic review is to answer two research questions: Q1: What are the trends of opinion mining and sentiment analysis publication from 2000-2016? Q2: Which are the areas in which opinion mining and sentiment analysis have been applied? Q3: What are the sources of data in the areas in which opinion mining and sentiment analysis have been applied? B. LOCATING STUDIES The objective of identifying an appropriate journal articles is to develop a list of all related articles to our research questions. We have selected Web of Science as a core database. The study has only focused on peer-reviewed articles that were written in English and published in the Web of Science Journal category. The study did not consider articles in other categories like conference proceeding papers, book chapters, review papers, and theses. Since the study is based on opinion mining and sentiment analysis, we used different strings to identify relevant papers. The researcher employed different techniques for searching, including separate keywords for ‘‘sentiment analysis’’ and ‘‘opinion mining’’, combining two keywords at the same time through simple operators and Boolean logic. For example, we used the string ‘‘opinion mining’’ and ‘‘sentiment analysis’’ for exploring the papers containing the exact phrases ‘‘opinion mining’’ and ‘‘sentiment analysis.’’ C. STUDY SELECTION AND EVALUATION In order to ensure and maintain the quality of the paper, we have constrained our selection of articles to only peerreviewed journals. Peer-reviewed journals have strict quality control and have gone through systematic, rigorous processes and have stringent requirements for publication, which leads to better research output [35]. The process began with scanning of selected articles from the Web of Science database. The timeline we set was from the years 2000 to 2016. The initial criterion of selection was based on choosing the keywords ‘‘opinion mining’’ and ‘‘sentiment analysis.’’ We have further followed the criteria suggested by [37] for the inclusion and exclusion of the articles from the selected list. Basically, we have developed the inclusion criterion as follows: • Published in peer-reviewed journals • Within the database of the last 15 years (2001 to 2016). • In the English language • Containing at least one keyword The exclusion criteria are as follows: • Have very narrow horizon or context • Do not explicitly focus on application side (human or organizational level) of opinion mining and sentiment analysis Our search resulted in an initial list of 274 articles for sentiment analysis and 91 articles for opinion mining. Thus, the preliminary results provided a total of 365 articles that were written on opinion mining and sentiment analysis. The next step involved reading the abstracts to evaluate whether it was relevant to our research topic. Initially a single person read it, but to warrant its rigor, an independent person to improve its objectivity and validity read the same number of articles. Scholarly outputs that did not align with our developed research questions or that seemed irrelevant and non-substantive were excluded. The articles that were included exhibited good fit with the objective of the study. Thus, a total of 99 articles were shortlisted based on the initial evaluation. In the next step, two authors reviewed the pre-selected articles separately. The total number of peer-reviewed journal papers selected for critical reviews after rigorous assessment were 58, published over a period of 16 years. The selected papers were then examined in detail and synthesized to answer the research questions. For the selected papers, we created the taxonomy which is represented in following sections (dataset, methods, application, and major challenges). IV. DATASETS A more in-depth analysis was done regarding the sources of datasets which are shown in Figure 2. The main source of data is Twitter with seventeen (17) articles, followed by movie reviews with eight (8) articles, Amazon with six (6) articles, blogs with five (5) articles, media with four (4), YouTube with three (3), and Tripadvisor with three (3) articles. There are also nineteen (19) articles that can be categorized as others with one (1) article from each. V. METHODS In this section we discuss the general process for sentiment analysis. A common process of sentiment analysis loop starts 37810 VOLUME 6, 2018 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges FIGURE 2. Opinion mining and sentiment analysis data sets. FIGURE 3. General process for sentiment analysis. with goal setting for employing the sentiment analysis. This depends on the application of sentiment analysis (see application section), and then big data is extracted either from a single or multiple sources. The next step is the application of specific sentiment analysis methods in order to mine this huge data and get insight into the company in making final decisions regarding their products. The general sentiment analysis process is shown in Figure 3. In the following subsection, we discuss the most commonly used method to analyze the extracted big data for sentiment analysis. First of all, the main process originates from the collection of the data from data sources such as social media. It should be kept in mind that the collected data should be relevant to the objective of the sentiment system. For example, if the company wants to collect their products, then the data should be extracted with keywords that represent the products. This also involves VOLUME 6, 2018 37811 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges setting the objective of the sentiment analysis to explain the purpose of sentiment analysis in order to determine the related keywords. The next stage is to pre-process the data in order to remove noise and unrelated content. It should be followed by the construction and evaluation of the sentiment analysis model (based on keywords, lexicon, or machine learning methods). As shown in Figure 4, machine learning algorithms methods commonly involve supervised machine learning methods which require training data to train the algorithms. After building the sentiments and evaluating them in real data sets (test data), the constructed model is applied to unknown big data to automatically classify the methods. Sentiment analysis models (based on keywords method, lexicon or based on machine learning method) are explained in the following subsections. A. KEYWORD-BASED CLASSIFICATION This method classifies text based on the presence of positive or negative polarity words such as happy, joyful, delighted, miserable, sad, terrified, and uninterested [38]. The main drawback of keyword-based classification is the inability to steadfastly classify the negated words and polarity, as this approach depends on surface features [38]. Another drawback is that this approach is based on the obvious presence of positive or negative polarity. However, occasionally, a post may covey sentiment or opinion through underlying meaning rather that obvious polarity words [38]. B. LEXICON-BASED CLASSIFICATION Lexicon-based approaches construct lists of words manually labeled as having positive and negative polarity, and a polarity score for each word is created. This constructed lexicon is used to calculate the overall sentiment score of a given post or text. The notable advantage of the lexiconbased method is that these methods do not need training data (as the supervised machine-learning method does). The lexicon-based method is widely used in conventional text like reviews, forums, and blogs [39], [40]. However, they are less likely to be used for big data extracted from social media websites [40]. The key reason is the unstructured format and nature of social media websites (the data contains textual peculiarities, informal and dynamic nature of language, new slang, abbreviations, and new expressions) [40]. Even though this approach outperforms the keyword-based classification, it still has drawbacks. Since it works at the word level, negated posts and posts with other meanings trick the lexicon polarity score measurement [38]. Second, lexical dictionary and polarity scores are usually biased toward the text of a specific type, dictated by the linguistic corpora source [38]. Therefore, it is challenging to construct a more generalized model regardless of the application domain. C. MACHINE LEARNING-BASED APPROACH Machine learning research has become a significant task in numerous application areas. Machine learning reaches throughout recent years have magnificently created algorithms for handling volumes of data to unravel realworld issues. Machine learning algorithms are grouped into supervised learning and unsupervised learning algorithms. Supervised learning algorithms will help users train and learn from the training example which is then tested and evaluated using the test data. The main drawback of supervised machine learning algorithms is the obligation to create a training example. The training example must be comprehensive enough to make the algorithm effective and reliable enough to classify the instance in test data. Another type of machine learning is unsupervised learning algorithms. The working principle of this algorithm is to identify the hidden associations in unlabeled data. The unsupervised learning methods are based on calculating similarity differences between data. For example, it calculates k-means in which similarity between data is computed based on proximity measures, such as Euclidean distance. The constructing machine learning-based method involves the following important steps. 1) FEATURES EXTRACTION As shown in Figure 4, feature extraction is an important part of building an effective machine learning method in which the textual posts (P1, P2, P3, . . . , Pn) are transformed into valuable word features (wf 1,wf 2,wf 3, . . .wfn) by using various feature engineering approaches. Feature extracting is one the most important steps of constructing effective classifiers [41]–[44]. The accomplishment or failure of the sentiment classification model is intensely dependent on the features quality. If the extracted features relate well with the sentiment polarity and can provide discrimination power between positive and negative, then classification will be more precise. In contrast, if the extracted features do not relate well with the sentiment polarity and the similar features exist in both positive and negative posts, then the classification task will be more challenging and less precise. The most commonly used features are the first automatic generated features of Bag of Words (BoW), Bag of Phrases (BoP), n-gram, and Bag of Concepts (BoC). The second group of features is based on lexical features such as opinion words, sentiment words, and negation words. The third group of features is varied based on the data source; for example, the data from social media normally adds features such as the number of hashtags and social media-related features such as abbreviations and emojis. 2) MACHINE LEARNING ALGORITHMS This subsection briefly describes the most commonly used machine learning algorithms in literature for sentiment analysis. a: ARTIFICIAL NEURAL NETWORK (ANN) The Artificial Neural Network (ANN) is a mathematical modeling approach that is stimulated by the operative processes of the human mind [45]. It is based on the artificial adaptive system which has some form of distributive architecture [46]. 37812 VOLUME 6, 2018 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges FIGURE 4. Steps for constructing machine learning based method. The system in ANN encompasses closely knit adaptive processing elements called artificial neurons or nodes which are proficient in carrying out enormous analogous and parallel computations for the purpose of information processing and knowledge representation [47]. These systems can also adapt to their internal structure in relation to a functional purpose. There are several types of artificial neural networks which have been used for different problems. Generally, their application is appropriate for a problem which is nonlinear in nature [47]. The nature of the problem can be carrying out pattern recognition, modeling memory, and envisaging the development of a dynamic system [46]. In most cases, these network types perform data modeling which is driven in a supervised or unsupervised fashion. The supervised learning technique is provided with both input and output. Then the network uses this input to generate output, and hence it is compared to desired outputs. On the other hand, in unsupervised training, only input is being provided and the network is used to find natural grouping with a dataset independent of external constraints. That is, the system itself must have the capability to decide which features to incorporate in order to group the input data. b: RANDOM FOREST Random Forests is a classification and regression method based on the ensemble of a proliferation of decision trees [48]. Recently, attention has been given to ensemble learning, a method which can create several classifiers and produce aggregate results [49]. The two commonly known and used methods for the classification of trees are boosting as proposed by [50] and bagging as proposed by [51]. The latter proposed the general operating mechanism of the Random Forest (RF), which augments the additional layer of randomness to bagging. In RF, each tree is a standard Classification or Regression Tree (CART) that uses the so-called splitting criteria like Decrease of Gini Impurity (DGI). Moreover, it picks the splitting predictor from a randomly chosen subset of predictors. Each tree is fabricated by using a bootstrap sample of the data, and the prediction of all trees are ultimately accumulated by means of majority voting [48]. VOLUME 6, 2018 37813 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges c: SUPPORT VECTOR MACHINE Support vector machines are considered to be universal learners. Generically they learn linear threshold functions. However, with the use of a suitable kernel function plug in, they can be used to learn in different applications in the form of radical basic function and sigmoid neural networks and be trained on polynomial classifiers [52]. SVMs were initially intended for binary classification, but research has extended it into a multiclass classification. There are two commonly used approaches for multiclass SVMs. The first deals with fabricating and conjoining the number of binary classifiers, and the second one directly involves keeping all data in one optimization construction [53]. d: GENETIC ALGORITHM The idea of developing the Genetic Algorithm was initiated and developed by John Holland. He proposed this idea in the year 1975 in his book ‘‘Adaptation in natural and artificial systems’’. Since then, the Genetic Algorithm (GA) has been increasingly recognized as a popular evolutionary computational research technique [54]. It has gained popularity over the years as an optimization tool in a variety of research domains, including computer science, operational research, engineering, management, and social sciences [55]. A major reason behind the realization of this technique is its diversified applicability, efficiency of operations, and applicability on global scenario [56]. Genetic Algorithms are search and optimization techniques in a multifaceted search space. They are inspired by the concept of genetics and natural selection. Some essential and principal ideas are adopted from the field of genetics and then used artificially to create a kind of algorithm which is flexible, robust, and efficient in nature [57]. Moreover, it characterizes the emergent technique which is to be used to understand different relationships in the course of the development of data, in which the data can come in the form of binary strings, formula, program, query, grammar, and images [55]. e: NAÏVE BAYES The Naïve Bayes (NB) classifier has long been used in most applications of supervised machine learning. It is considered a tool for the retrieval of data [58]. It is based on a simple theorem of probability for making a probabilistic model of data. The mechanics of the NB algorithm are applied to numeric data [59]. It is simple, easy to understand, and quick for classification. It normally entails a minimal data set for training and then is used to predict the parameters needed for classification purposes. f: DECISION TREE The decision tree (DT) classifier has been widely used for prediction and classification of tasks. The rules in creating the decision tree are easy to understand. The classifiers built through the decision tree are given in hierarchical representation. The tree is composed of decision nodes, event nodes, edge, and path [60]. A variety of classifiers are used in a variety of applications. Some of the DT classifiers are ID3, C4.5, and C5.0, but a common problem that has been found in the DT classifier is the ability to incorporate all types of variations in data, including noise when trees get bigger and deeper. This problem is commonly known as overfitting. In addition, the structure of the tree is distorted with the addition of data. To avoid this problem, the random forest technique is recommended in which many trees are formed and trained by dividing the training sets, and outcomes are generated through aggregation of all trees. g: K-NEAREST NEIGHBOUR The K–nearest neighbor algorithm performs classification based on instance learning. It works through a non-parametric procedure of storing all inputs and instances and classifies new inputs by means of similarity measures like Euclidean distance [54], [55]. h: ENSEMBLE VOTED CLASSIFIER An ensemble of classifiers is the collection of many classifiers in which their decisions are aggregated by means of weighted an unweighted voting mechanism to predict the outcome. The classification in the ensemble voted classifier is based on a voting mechanism in which classification of new instances is done by considering the majority vote of the prediction [51], [56]–[58]. D. COMPARISON BETWEEN THE SENTIMENT ANALYSIS METHODS Table 1 summarizes the various common methods used in the sentiment analysis literature. The extant literature shows the effectiveness of the machine learning algorithm. However, with the increase in its effectiveness, the complexity of the method has also increased. Table 2 shows a comparison between the working principle, advantage, disadvantage, and time complexity of the most commonly used machine learning algorithms for sentiment analysis. VI. APPLICATION AREAS OF OPINION MINING AND SENTIMENT ANALYSIS Opinion Mining and Sentiment Analysis has gained popularity in recent years and has been applied in many application areas. It has been used in diversified areas like healthcare, the financial sector, sports, politics, hospitality and tourism, and consumer behavior, as highlighted in Appendix 1. Some of the growing and emerging application areas will be reviewed in this section. A. HEALTHCARE Korkontzelos et al. [70] enhanced the state-of-the-art adverse drug reaction method (ADR) by incorporating sentiment analysis features. The purpose is to examine the influence of sentiment analysis features in locating ADR mentions. For this, annotated ADR posts related to 81 drugs are collected from a Twitter and DailyStrength forum. Then, sentiment 37814 VOLUME 6, 2018 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges TABLE 1. Summary of the common methods used in sentiment analysis. analysis features are added to evaluate their effectiveness in locating ADR mentions. The results show that sentiment analysis features slightly boost the performance of ADR mention in both tweets and health-related posts in the forum, which can be utilized as pharmacovigilance practice in the future. Rodrigues et al. [71] studied the mood of cancer patients by using a sentiment analysis tool named SentiHealth-Cancer. The aim of the study is to uncover the emotional condition of cancer patients though their discussion in the Portuguese language among the Brazilian online community. The posts are collected from the two different cancer communities on VOLUME 6, 2018 37815 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges TABLE 2. Comparison of working principle, advantage, disadvantage, and time complexity of most commonly used machine learning algorithms for sentiment analysis. 37816 VOLUME 6, 2018 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges Facebook. The proposed tool, SentiHealth-Cancer (SHC-pt), is compared with AlchemyAPI 1.1.4v, Seman-tria 3.0.67v, SentiStrength 0.1v, and Textalytics 1.2 v in six different experiment settings. The result shows that this tool outperformed other sentiment analysis in all experiments including the complex setting, which shows that this tool is far better than other tools, especially in the cancer context. Kim et al. [72] examined the coverage and sentiment trends of different media sources. The study relied on two sources: news publications and Twitter. The aim of the study is to evaluate the difference in coverage on issues like the Ebola virus and investigate the difference in sentiments on both mediums and to check the degree of change of sentiments over different time periods. For the experiment, about 16,189 news articles from around 1006 different publication sources and 7,106,297 tweets were collected. The experiment revealed that both Twitter and news media are different mediums of communication in terms of coverage and sentiment dynamics. The experiment on Twitter indicated that it is a timebound medium and seems to be narrower in terms of coverage and have a shorter life span than news media. B. FINANCIAL SECTOR Hai et al. [73] developed a stock price prediction model by incorporating sentiments of specific topics related to the company. Two datasets, including historical price dataset and mood information dataset, have been used to evaluate the effectiveness of the model. Historical stock prices of 18 companies extracted from Yahoo Finance and messages related to stock prediction, mood of the investors, discussion related to management of the organization, and specific events were extracted from the Yahoo Finance message board. The support vector machine (SVM) was used as a classifier, and six features were applied, which include price, human sentiment, sentiment classification, Latent Dirichlet Allocation (LDA) based method, joint sentiment/topic (JST) based method and aspect-based sentiment were designed and incorporated to evaluate the effectiveness of sentiment analysis in predicting stock market movement. The result shows that the performance accuracy of the proposed sentiment drove stock price model is approximately 2 percent better than the model, which only uses historical prices. Moreover, accuracy in predicting difficult stocks is approximately 10 percent better than historical price driven method. Li and Meesad [74] also worked on predicting stock market trends, but the study focused on reducing socialization biasness by proposing a model which blends both sentiment analysis and socialization biasness through inverse bias algorithm. The proposed model is based on the support vector machine with hybrid features. For the experiment, 4,622 tweets were collected from Topsy.com during the timeframe from July 1, 2013 to May 30, 2014. The result indicated that SVM functioned better than other algorithms like Naïve Bayes and K-nn. Moreover, the model based on linear SVM classification with hybrid feature selection has improved its accuracy (from 84.06% to 86.95%) in predicting stock market movement when using 10 – folder cross validation. In contrast the accuracy of inverse socialization bias also increased from 86.95% to 90.33% while incorporating with linear SVM algorithm. Chen et al. [75] investigated the relationship between public opinions or emotions with the changes in the stock market of China. The experimental dataset used for this study is Weibo, a social media website. Initially, the Chinese segmentation tool Jieba, which is based on dynamic programming, is used to fragment the text into different words and then categorize them into seven basic emotion categories which is based on Chinese Emotion Word Ontology. The result based on Granger causality analysis suggested that emotional states like happiness and disgust are strongly predictive to stock market prices in China. Moreover, the study demonstrated that, with the use of basic neural network model, emotional states aids to predict movement in the stock prices. C. SPORTS Schumaker et al. [76] studied on prediction of English Premier League matches outcomes through a central sport system based on the crowdsourced sentiment technique. Moreover, the system also used it for wagering decision purposes. For the experiment, twenty clubs were selected, and the tweets from the last three months of the English Premier League season were analyzed. During that time, 122 matches were played, and about 96 hours of tweets containing the club hashtag before each match were collected from the Twitter API. During that time period, 18,027,966 tweets were collected. The results suggested that crowdsourced sentiment outperformed in predicting the match outcome than the crowdsourced odds. The analysis suggests that tweet sentiment outshines odd favorite wagering and has better payout returns in comparison to non-favorite wagering. Furthermore, goal difference and net payout return are improved with overwhelming positive sentiment. The analysis of the study proposes that professional odds markedly forecast non-positive match outcomes and show close goal margins. Yu and Wang [77] studied sentiment analysis on tweets of US sports fans during five games of the FIFA World Cup 2014. The objective of the study was to evaluate their emotional responses and changes during the games, especially after their own or opponents’ goals. For this purpose, several tweets were collected from three US games and two non-US games for comparative analysis. The results suggested that fear and anger were the dominant emotional state that US fans expressed when the opposing team scored, and that emotional state decreased with the US team goal result. Moreover, emotional responses of US fans in nonUS games were ambiguous, but surprisingly, fans’ tweets are more inclined towards positive emotions like joy and anticipation than negative emotions. The result of the study showed consistency with predictions of disposition theory, which states that emotions can be more predictable with the clarity of the fan base. VOLUME 6, 2018 37817 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges D. POLITICS Ceron et al. [78] used social media forums to investigate and follow the political preferences of the general public. The study focused on examining the pattern of the online popularity of different Italian political leaders throughout the year of 2011 and analyzed the intention of French voters in choosing the candidate for French Presidential ballot in 2012 and then following legislative elections. The study employed Hopkins and King’s method of analyzing tweets. The result showed that social media platforms have a great capacity to predict electoral outcomes, and the result is highly correlated with old-fashioned methods like mass surveys. In addition, this study also emphasized and elucidated the increasing power of predictability through social media data analytics with the increase of opinions and expressions online. Alfaro et al. [79] worked on opinion mining and sentiment analysis by integrating supervised and unsupervised machine learning techniques through a multi-stage procedure for the purpose of automatic recognition of various opinion trends in the weblogs. The study employed faculty and university weblog comments to test the framework. The test result indicated that SVM classification technique outperformed k-NN in terms of accuracy; however, integrating both approaches may increase accuracy. Based on the result, Alfaro et al. [79] anticipated that the proposed technique and procedure may be applied in different domains from the electoral campaign to develop a public policy or new law. Furthermore, it can be applied in analyzing feedback on a company’s products and services and can be linked to the company market campaigning activities. E. HOSPITALITY AND TOURISM Philander and Zhong [80] studied sentiment analysis on Las Vegas resorts as a case study. The study employed Twitter data to reveal the application of sentiment analysis as low cost and a real-time tool in evaluating customer perceptions about the services. Using Twitter data, a sentiment index was created using a sentiment lexicon methodology. The resulting sentiment metrics were used for performance comparative analysis of different firms over different time periods. The outcome of the sentiment score is then compared with data from TripAdvisor to evaluate external validity, which shows that both sentiment metrics and TripAdvisor are very much the same in terms of convergent and discriminant validity. The analysis shows that Twitter contains broad, direct, and indirect perspective about people’s opinions towards Las Vegas properties, unlike TripAdvisor, which mainly focuses on the hotel customers’ experiences and perception of facilities and services. Rodolfo et al. [81] utilized text comments with the aim of applying sentiment analysis through three different algorithms in predicting overall hotel rankings. For this purpose, the numerical ratings and more than a million reviews of hotels situated in seven cities were gathered from the TripAdvsior web portal. Reviews were deemed as positives and negatives by using three different sentiment analysis tools. The results indicated that all classifiers are positively correlated with the actual rating of TripAdvisor to support the argument that the textual data of users can be transformed into numerical ratings. Moreover, the reliability of all three classifiers in predicting the hotel rating were compared with actual ratings and found complex algorithms, which, based on boosting and recursive neural tensor networks, were better in performance in relation to simple Naïve Bayesian algorithm. The study would be useful in areas like traveller forums by simply summarizing and consolidating opinions of potential and existing customers and can be used for predicting the rating. F. OTHER APPLICATION AREAS Chung and Zeng [82] developed a framework for policy makers and evaluated public opinions about US immigration and border security on social media through a system called iMood based on sentiment and network analysis. For the experiment, around 909,0350 tweets were evaluated on the grounds of 300,000 users’ sentiments and emotions in three different phases starting from May 2013 and ending in November 2013. Based on the analysis, the study highlighted significant changes in emotion and sentiment among different time zones or phases. The study would be a starting point in framing a policy in relation to public sentiments. Liang et al. [83] employed a multifaceted sentiment analysis in predicting the sales of mobile apps based on online word-of-mouth textual comments. The comments related to product quality and service quality of 79 paid and 70 free apps are extracted from the iOS store. The test indicated that product and service quality are the dominant features in the eyes of the user and have a significantly positive impact on the ranking of app sales. Kim et al. [84] mined public opinion from a social media networking platform in predicting box office performance based on a movie trailer. For the experiment, comments or reviews of trailers for 200 different movies were gathered from YouTube to analyze sentiments about movies. Moreover, the marketing property data of the chosen movies were collected from IMDB and the box office website, and these movies were categorized into different classes based on business. The results suggested that combining viewer comments and marketing properties assisted in improving prediction about box office performance. Zhou et al. [85] exploited multi-granularity sentiment analysis approach in understanding the behaviors of consumers in different regions. The study conducted on Chinese and American customers of digital camera, smart phones, and tablet computers. The study revealed interesting findings, including that American customers are blunt and generally direct in expressing their views about the product features. On the other hand, Chinese customers most of the time used soft expressions. Moreover, American customers emphasized product details and internal features more, while the Chinese paid more attention to general feelings and the external features of products. 37818 VOLUME 6, 2018 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges D’Avanzo and Pilato incorporated a cognitive approach based on collaborative learning to improve buyers’ online shopping decisions. The proposed approach is experimented on Facebook reviews of two markets: the smart phone brand Nokia and the fashion brand Zalando and Privalia. The study employed Bayesian social sentiment technique and summarized opinions from different markets in order to arrive at a specific decision. This would help speed up the shopping activity and ultimately enhance the online shopping decision. The authors claimed that manufacturers or sellers would benefit through this approach, either by improving their offerings to the buyers or changing the product lines based on customers’ feedback. VII. OPEN CHALLENGES In this section we discuss the open challenge of big data sentiment analysis. We discuss both technical (challenges related to the development of Sentiment technique) and nontechnical challenges (challenges related to the application of Sentiment analysis). A. CHALLENGES RELATED TO THE DEVELOPMENT OF SENTIMENT TECHNIQUE In this section, we discussed the challenges related to the development of sentiment and opinion classifiers. The following section will discuss first the heterogeneous characteristics of big data-related challenges and the future directions needed to develop a method that can effectively deal with the heterogeneous characteristics of big data. The second challenge related to users’ network (future researches need to further investigation on how to make use of user’ relationship network for enriching the output of the sentiment system). Third challenge related to analyzing sparse, uncertain, and incomplete data and future researches are suggested to focus on developing sentient analysis method which can effectively be able to handle sparse, uncertain, and incomplete data. Lastly Semantic relations in multi-source data fusion. 1) HETEROGENEOUS CHARACTERISTICS OF BIG DATA Big Data occupies several characteristics but massive amount of data in heterogeneous and diverse formats (unstructured data) are the notable characteristics of big data. Such characteristic is created due to diverse strategies of how the data is collected, from where the data is collected and nature of diverse applications. Considering this characteristics, the sentiment classifier should be work effectively with such heterogeneous characteristic of big data. Usually traditional sentiment classifier was dealing with data from one source for example the company online review or company feedback records. However, in the era of big data, the sentiment classifiers should have the ability to handle diverse data from different data sources. Big users’ network, In most above discuss studies have analyzed the post in order to classify the sentiment of the post regardless how an opinion is formed. However, with introduction of with big data such as big social data from social media website such twitter, rich information of people and their network with such platform can drive the researchers into deeper analysis of how the sentiment is formed. Individuals form similar friend clusters constructed according to on their common hobbies, views, interests, family relation or geographical relation. Such social contacts usually occur in not only our everyday activities, but similarly remain very common in virtual worlds [86]. The impact of users networks in forming the sentiments need more Investigation which may hold significance important strategies for smarter marketing method which does not only understand the opinion polarity of a post but also understands how this post is constructed within the social connections. 2) ANALYZING SPARSE, UNCERTAIN, AND INCOMPLETE DATA One of the characteristic of big data is that it contains a lot of noise because of the wide use of abbreviations and misspellings. This phenomenon, known as data sparsity [40], has an effect on the accuracy of the sentiment classification. Moreover, the big data (for example social media sites) are incomplete due to the privacy restriction. Therefore, some information such as location that can be used as features to enhance the functionality of sentiment classifier. For example, building sentiment analysis of political posts within specific country needs to understand the location of the posts. However, geolocation information may exist in some posts and not in other posts. Therefore, future works are required to create sentiment classifiers which should be able not only to classify the sentiment polarity of a message but also to predict the incomplete information to provide more accurate details to construct a complete application at the end. 3) SEMANTIC RELATIONS IN MULTI-SOURCE DATA FUSION Twitter, Facebook, Instagram, and YouTube may discuss an event at the same time. Analyzing semantic relations in these data sources can offer better insights and better understanding of whole sentiment picture. Analysis of an event from different sources [86] and then constructing semantic associations between text, image, and video data will significantly improve and enrich the output of the sentiment analysis systems [86]. Nevertheless, it is a challenge to construct such a semantic association based on models to fill the semantic gap between such heterogeneous data sources [86]. This can introduce a great opportunity to create sentiment analysis models based on multi-source data fusion. B. CHALLENGES RELATED TO THE APPLICATION OF SENTIMENT ANALYSIS 1) DOES SENTIMENT ANALYSIS HELP IN DESIGNING ENTERPRISE STRATEGIES? Big data has become an important element for enterprises to understand customers’ opinions about their products. Big data sources such as social media provides huge usergenerated data which is a worthy source of opinions and are treasured for many applications that involve comprehension VOLUME 6, 2018 37819 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges Table 3. Synthesized areas of opinion mining and sentiment analysis applications. 37820 VOLUME 6, 2018 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges Table 3. (Continued.) Synthesized areas of opinion mining and sentiment analysis applications. VOLUME 6, 2018 37821 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges Table 3. (Continued.) Synthesized areas of opinion mining and sentiment analysis applications. 37822 VOLUME 6, 2018 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges Table 3. (Continued.) Synthesized areas of opinion mining and sentiment analysis applications. of the public opinion about events, products, persons, etc. For example, an enterprise that may capture the opinions of clients about their products [40] may apply some sentiment analysis to advance the quality of their products. However, more investigation is required to understand more factors that can work together with sentiment factors to better understanding the public opinion. For example, product ‘‘A’’ has many positive sentiments in the recent three months, but does it mean that the enterprises which produced product ‘‘A’’ need to maintain the quality of this product, as it still has a positive sentiment by the users, and does the enterprise need to increase the production of this product to meet the popularity of the product? Sentiment analysis alone may not be enough to answer these questions. Therefore, factors like competitors’ information, country economic growth, consumer confidence index, and other factors may influence conclusive decisions. However, future studies need to integrate sentiment analysis methods with other factors to create effective methods for a fully automated decision-making system based on several factors. Moreover, future research needs to investigate whether big data-based methods can replace traditional methods (questionnaire-based methods) in making decisions or whether both can work best when integrated into a system. 2) THE INFLUENCE OF THE POST A user (A) with a large number of followers posts a negative comment about product X, and a user (B) with few followers posts a positive comment about the same product. Statistically, we have same number of positive and negative comments of product X, but in reality, the negative post should have more impact, since user A has several followers who can be influenced by his/her opinion compared to user B, who has few followers; consequently, his opinion influences fewer users compared to user A. Moreover, in many cases, the direct number of followers does not indicate high influence [83], [84]. Therefore, future research should further investigate the influence of sentiment polarity within the connected networks of users. Integrated sentiment analysis and influence measurement can provide a precise sentiment measurement which takes into consideration both the polarity of the post as well as its influence to capture consumers’ views on a larger and deeper scale. 3) THE IMPACT OF SOCIAL BOTS Recently social bots have become sophisticated as well as threatening. The presence of social bots, especially in social media, can be a threat to online systems as well as to our society [89]. Social bots are software programs aimed to simulate human user on social media websites. Gradually, politicians, marketer, and different firms use these automatic social bots in online environments in order to manipulate public opinion [90]. For example, political bot-based strategies are used to enormously increase politicians’ followers on social media sites and generate positive comments to create impressions of popularity [90]. Similarly, this method can be applied in other areas as well. Therefore, two questions may arise: first, is the output of sentiment analysis based on the genuine opinion of the user or is it a bot-generated opinion? Second, to which extent can the social bot influence public opinion? Hence, in future research, these two questions must be investigated in order to know how the opinion is formed and to recognize how sentiment analysis should be designed to take this factor into consideration. VIII. DISCUSSION AND CONCLUSION A few conclusions can be made based on the findings of the systematic literature have been conducted on articles published in the Web of Science from 2000-2016 on OMSA. In addition, more articles have been published on sentiment analysis as compared to opinion mining since 2015. The emergence of the significance of sentiment analysis matches the development of social media usage, including reviews, forum blogs, micro-blogs, Facebook, Twitter, and other social networks. Interestingly, presently we have access to a huge amount of opinionated data which can further be used for different methods of analysis. Typically more than 80% of social media data can be monitored for analysis purposes [135]. VOLUME 6, 2018 37823 S. Shayaa et al.: Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges For instance, a tweet contains a maximum of 140 characters, therefore monitoring software can assign a specific sentiment score to that tweet. That sentiment score represents a semantic judgment to examine whether the tweet seems to be positive, negative, or neutral. Gathering reviews on products and services on the other hand is mainly done using opinion mining. Products and services are considered as entities, and the process of mining opinions usually involves performing opinions of the texts. The trends further indicate that society is more likely to express feelings rather than what they think on social media platforms. In terms of applications, more research has been done on the assessment or evaluation of the various methods of opinion mining and sentiment analysis. Although these refer to the evaluation of the techniques used, the data sets extracted from users’ application databases thus include an element of human application. Marketing-related activities still dominate the applications followed by the financial, healthcare, and hospitality and tourism industries. It is further noted that applications of opinion mining and sentiment analysis for politics and government views are still growing. The data generated by people through their views on current events in the country can be very useful to political parties and for the general public in resurrecting the future course of action for their interests. In the datasets used for OMSA, data from Twitter seemed to dominate the data sets. This is further aligned with the growth of the sentiment analysis in which most of the data is captured from social media more profoundly on Twitter as compared to the other data sources. Previously, Twitter has been used as a tool for disseminating and propagating information rather than simply a social networking site. Previous research shows that top users, as measured by the number of followers on Twitter, are mostly celebrities and those who attract the keen interest of the mass media [91]. An electronic platform for word-of-mouth influence in marketing [92], Twitter also serves as a political sentiment analysis predictor of elections [93] and as a stock market movement predictor [94]. This has widened its exploitation to politicians and other pundits. It can be used to quickly share information with people, promote new products, and communicate with celebrities’ fans or political supporters.