Sentiment Analysis Based on MapReduce: A survey

Conference Paper · December 2018

DOI: 10.1145/3291280.3291795

3 authors:

Mariam Khader
Princess Sumaya University for Technology
18 PUBLICATIONS **24** CITATIONS
SEE PROFILE

Arafat Awajan
Princess Sumaya University for Technology
99 PUBLICATIONS **367** CITATIONS
SEE PROFILE

Ghazi Al-Naymat
Ajman University
61 PUBLICATIONS **512** CITATIONS
SEE PROFILE

Sentiment Analysis Based on MapReduce: A survey

Mariam Khader,
Computer Science, Princess
Sumaya University for Technology
Amman, Jordan
mkhader@psut.edu.jo

Arafat Awajan
Computer Science, Princess
Sumaya University for Technology
Amman, Jordan
awajan @psut.edu.jo

Ghazi Al-Naymat
Computer Science, Princess
Sumaya University for Technology
Amman, Jordan
g.naymat@psut.edu.jo

ABSTRACT

Sentiment analysis is the process of analyzing people's sentiments, opinions, evaluations and emotions by studying their written text. It attracts the interest of many researchers, since it is useful for many applications, ranging from decision making to product evaluation to mention a few. Sentiment analysis can be conducted using machine-learning techniques, lexicon-based techniques or hybrid techniques that combines both. As people are more reliant on social networks such as Twitter, this has become a valuable source for sentiment analysis. However, the existence of big data frameworks require adaptation of these techniques to run within such frameworks. This paper reviews sentiment analysis techniques, focusing on the MapReduce-based analysis techniques. We found that the Naïve Bayes algorithm was the most used machine learning technique for extracting sentiments from big datasets because of its high accuracy rates. However, the dictionary-based techniques achieved better results in terms of execution time.

KEYWORDS

Big Data, Dictionary Based Analysis, Machine Learning, MapReduce Framework, Naïve Bayes, Sentiment Analysis.

1 Introduction

Big data is referred to large and complex data sets that cannot be handled by traditional systems. The big data is defined by its three V's - volume, variety and velocity. The volume means a huge amount of data, variety is referred to different types of data and velocity is about the speed of data processing [1]. Main sources of big data include social networks. Mainly, with the increasing number of users and the uploaded contents. Until 2013, only 0.5% of the data in the universe has been analyzed and this proportion is decreasing as the production of data is exponentially increasing [2]. One of the techniques for analyzing large data sets is Hadoop-MapReduce. Hadoop is the most well-known big data frameworks, which is based on the MapReduce programming model and the Hadoop Distributed File System (HDFS). The HDFS is used for distributed storage of data, while MapReduce provides a distributed processing application for handling massive volume of data.

One of the most used applications to extract valuable information from data sets is the sentiment analysis, where a sentiment or an opinion is extracted from a text. The sentiment analysis defines the contextual polarity of a text using natural language processing (NLP) techniques. It can be achieved using machine-learning techniques, lexicon-based techniques or both [3].

People are using social networks as free environments for conveying their opinions, emotions and feelings about products, films, events, etc. Twitter is considered the most used social network for mining opinions [4]. Hundred million active user daily and 500 million tweets are sent every day. Each tweet must be less than or equal 140 characters. It can contain text, hashtag (#), other user mentions (@) and URLs. The increase number of tweeter's users and tweets published daily makes Twitter an attractive source to analyze people feelings and opinions. For example, in business development, it is essential to study the tweets and posts of social networks to understand the customers' opinion and feelings about a specific subject. In essence, the customers' opinion about a new item, a brand or product contributes in the business development.

As the size of data produced by social networks is rapidly increasing, single machine based sentiment analysis cannot handle the huge volume of data. Parallel frameworks provide better environment for large data analysis. Using MapReduce framework for parallel sentiment analysis has attracted many researchers because of its simplicity, scalability and fault tolerance.

Many research have utilized the MapReduce framework to extract sentiment analysis from big data sets. Thus, in this paper, the focus will be on MapReduce based sentiment analysis techniques, which are considered the most adequate framework for the parallelization of sentiment analysis. The paper highlights the main contribution of developed techniques, used algorithms, their efficiency, analysis steps and results.

The rest of this paper is structured as follows: Section 2 presents an introduction of the sentiment analysis concepts and steps. The literature review of traditional sentiment analysis techniques is discussed in Section 3. Section 4 presents the sentiment analysis techniques on the Hadoop MapReduce framework. Finally, Section 5 includes the conclusion remarks.

2 Sentiment Analysis

Sentiment analysis is defined as the process of applying NLP and text analysis methods to identify and extract information from unstructured text. The sentiment analysis is employed in many applications [5]. It supports the decision-making, where extracting peoples' opinions from their comments and reviews helps to make the adequate decision. For instance, many companies produce new products or improve their current ones based on peoples' reviews. It is also utilized to determine expectations in elections based on people sentiments about candidates.

The following subsections discuss important concepts related to sentiment analysis.

2.1 Subjectivity/Objectivity

In order to apply sentiment analysis, objective and subjective texts should be identified. Since only subjective texts hold the indication of the sentiment. While objective texts include information. For instance, the sentence: "An extremely epic and well-done film" contains a sentiment (well-done), therefore it is subjective. On the other hand, the sentence: "Braveheart movie was directed by Mel Gibson" implies a fact, does not contain any sentiment, therefore, it is objective.

2.2 Polarity

After identifying subjective texts, its polarity is examined. Text polarity is categorized into one of the following:

2.2.1 Positive: positive text holds a positive sentiment. For example, "The end of the story was brilliant". This sentence represent a pleasant feeling in the users' opinion.

2.2.2 Negative: negative text holds a negative sentiment. For example, "The historical facts are inaccurate; I cannot believe people actually believe this stuff happened." This sentence represents unpleasant feeling in the users' opinion.

2.2.3 Neutral*:* neutral text does not hold a positive or negative sentiment. For example, "I usually get hungry by noon." Such sentence contains a user's feeling, but it does not reflect a positive or negative polarity, therefore it is neutral.

These three categories, positive, negative and neutral determine the polarity of the text. Researches may consider two classes of text" positive/negative or the three classes. However, it has been found that considering the neutral class increases the accuracy of the sentiment decision. The classification using three classes is performed in two ways: the first is classifying the text in two phases, first into neutral or positive/negative and then handle the positive/negative class. The second is to classify the text into three classes in one phase.

2.3 Sentiment analysis levels

The text can be processed on different levels to determine its polarity.

2.3.1 Document Level: in this case, the input document by whole is classified into positive, negative or objective [5]. The document-level sentiment analysis is considered the most difficult level of analysis [6]. The challenge is that not all sentences in the document actually indicate a sentiment. Because of this, determining subjectivity/objectivity of each sentence becomes important to enhance the performance of the analysis task [7].

2.3.2 Sentence *Level:* in this case, each sentence is processed separately to determine its polarity.

2.3.2 Phrase *Level (Aspect-based analysis):* in this case, the document is processed more deeply, where phrases or aspects of the sentence is to determine its polarity.

2.4 Features selection

The features selection (or extraction) is important step in determining text polarity. These features can be:

2.4.1 N-grams: which represents a contiguous sequence of n terms from a given text. It can be "unigram", process one word at a time, or "bigram", process two words at a time, up to n words. The sentence "The movie seems good" For n=2 (bigrams) "this movie", "movie seems", "seems good". Not all sentiment can be realized using unigram features. For example, "I cannot hate such a movie". Using unigram analysis this sentence indicates negative sentiment (hate), however using bigram analysis it indicates a positive sentiment.

2.4.2 PoS-tagging*:* The part of Speech (PoS) is a way to determine the indication of a word in the text, with respect to its definition and its relation with contiguous words. A word is tagged as a Noun, an Adjective, an Adverb or a Pronoun, etc. Adjectives and adverbs are the most important tags in sentiment analysis as they hold most of the sentiments in the text.

2.4.3 Stemming*:* The stemming is the task of removing affixes (prefixes or suffixes) from the word. For instance, "watching", "watches" are stemmed to "watch". The stemming helps in sentiment classification but it is found that it sometimes might decrease the classification accuracy.

2.4.4 Stop words*:* Removing stop words from texts can provide a little information about the sentiment. Most of researcheres employ stop word removing as a preprocessing step. Examples of stop words are Pronouns (he, she and it), articles (a, an and the) and prepositions (in, near and beside).

2.4.5 Conjunction handling*:* Generally, sentences express one meaning at a time, but the existence of some conjunction words, such as however, although, but and while, may change the text meaning. For example, "The story of the movie was interesting; however some actors were not good as expected". It has been found that conjunction-handling can increases the accuracy.

2.4.6 Negation handling*:* negation by using words as "not, neither and never", inverts the polarity of the sentence. For example, "the plot was not good at all" indicates unpleasant feeling, even though good is a positive word.

2.5 Sentiment classification

Mainly, two approaches are used for text classification:

2.5.1 Subjective lexicon approach: The subjective lexicon is a collection of words, where each word is assigned by a score that determines how positive, negative neutral or objective is a text. This approach aggregates the score of words separately for positive, negative neutral and objective. The highest score between these four scores determines the polarity of the text. The lexicon approach is applied using Dictionary-based approach or Corpus-based approach.

2.5.1.1 Dictionary based approach. Using this approach, a set of sentiment words are gathered manually to form a seed list. Then the synonyms and antonyms of words are searched in the dictionary or thesaurus and added to the seed list.

2.5.1.2 Corpus based approach. The corpus is a collection of documents, usually on a specific subject. In this approach, the seed list is expanded using the corpus text. Thus, it helps in determining the orientation of the text. This approach is accomplished in two ways:

A. Statistical approach. It works by calculating the frequent words in a corpus, i.e., if a word appears mostly in a positive text, then it is assigned a positive polarity, but if it appears in negative text, then its polarity is negative.

B. Semantic approach. In this approach, the sentiment value is calculated by measuring the similarity between words, usually by finding synonyms and antonyms of a word. Most research utilizes the wordNet for this purpose.

2.5.2 Machine learning approach: the classification is automatically performed using machine learning on extracted features. Using machine learning techniques, a system model is built based on labeled data. Where each class represent several features and has a label assigned to it. The used classifiers can be probabilistic classifier, which predict the class of the text based on some features. An example of such classifiers are the Naïve Bayes, Bayesian network and Maximum entropy. Another classifier is the Linear classifier, which classify text based on linear combinations of the value of the characteristics. An example of this classifier is SVM and Neural network. The decision tree classification employs a condition to classify the data. The rule-based classifier employs rules in the training phase for classification decision [5].

3 Literature Review

The social networks introduced a rich source for peoples' emotions, opinions and attitudes, which can be utilized for the research of the sentiment analysis. Huge research exploits the social networks for sentiment analysis. It has been applied in different applications such as finding peoples' opinion about a product or new brand, predicting the results of elections or in e-learning system. In [8] an adaptive e-learning system was proposed. The system analyzed the students' personality to determine the students' motivation for studying a specific course. Based on the student twitter's profile, the system classified the students into motivate, demotivate or neutral. The results helped to determine whether to give the student a course or not based on these results. The authors utilized the AFINN Dictionary to classify the tweets. According to [9] [10], the methods of sentiment classification can be categorized into lexicon-based methods, machine-learning methods and hybrid methods. The lexicon-based methods depend on a sentiment lexicon, which is a set of predefined sentiment terms. It can be dictionary-based method or a corpus-based method, where semantic and statistical approaches are used to determine the sentiment polarity. The machine learning (ML) methods apply linguistic features and it employs known algorithms such as Naïve Bayes, SVM and Neural Networks. The ML methods are generally used to classify a given text as positive or negative based on an opinion or sentiment included in this text [11]. The hybrid analysis methods combine ML and lexicon-based methods. The main role in these methods is for the lexicon-based techniques. A hybrid approach was proposed in [12]. For computing the sentiment, the method applied basic techniques of NLP by combining the fuzzy sets into the SentiWordNet to determine the orientation polarity and the intensity for each sentence. In [13] . authors proposed a hybrid sentiment analysis approach for Arabic texts. The approach utilized the fuzzy logic and the lexicon-base techniques to determine the sentiment polarity. The proposed approach had two steps: First, the Arabic text was assigned weight. In the second step, the fuzzy logic was applied to assign the polarity for each sentence.

Several methods were proposed to enhance the quality of sentiment analysis results. It was proven by [14] that integrating the emoji characters in the text analysis can enhance the sentiment analysis results. The method was applied on a data set of positive and negative events, which was extracted from Twitter. Two methods for sentiment analysis were proposed by [15]. The methods were the "sentiment classification algorithm" (SCA), which was based on k-nearest neighbor (KNN), and the second was based on the support vector machine (SVM). The tweets were classified into positive and negative. The research of [16] emphasized the importance of techniques of text preprocessing in improving the sentiment classification accuracy. Six methods of text preprocessing was used in combination with two of feature models and four classifiers that were applied on a twitter data set. The results of the research concluded that applying acronyms expanding and negation replacement enhanced the results accuracy. However, the impact of removing URLs or removing numbers and stop words had a rarely impact on the results. The role of NLP data preprocessing on sentiment analysis was also studied by [17]. The used preprocessing techniques were stop words removal, HTML tags cleanup, negation handling, stemming and abbreviation expansion. The results indicated that the used presentation and features had improved the accuracy of sentiment classification.

In [18] a research applied stop words removing to enhance the sentiment analysis results. The accuracy was impacted negatively by using precompiled stop word lists. Five text-preprocessing methods were used in [19] to demonstrate the results of text preprocessing methods on sentiment classification. The five methods, which are stemming, negation, lemmatization, removing repeated litters and URLs were applied on a Twitter dataset. Among these methods, the features reservation, negation transformation, use of URLs, normalization and repeated letters enhanced the results accuracy. While stemming and lemmatization decreased the accuracy results. The research results indicated that augmenting the feature space with emotions features and bigram had enhanced the results. The effect of negation for sentiment analysis was studied in [20]. The research proposed a new approach for handling negation to enhance the results accuracy. The method, which was applied on a Twitter data set, showed promising results that enhanced the accuracy, Recall and Precision. Other research focused on adding the semantic analysis to enhance the accuracy of the sentiment classification. It was proven that employing semantic methods solved problems relate to NLP and improved the results of sentiment analysis. A new method was proposed by [21] to extract the words patterns of similar sentiments and contextual semantics in the tweets. The results of the proposed method was better than other approaches. [22] proposed a semantic method for determining users' sentiments from Arabic text. It introduced an Arabic Sentiment Ontology (ASO), which includes words that express feelings. The method has classified different topics correctly. [23] proposed a lexicon-based method for semantic polarity extraction using fuzzy logic. The method was tailored for Arabic texts and had two steps. First, the text was assigned weight. In the second step, the fuzzy method was used to determine the polarity of the sentences.

4 MapReduce based Sentiment Analysis

The characteristics of big data create new challenges into the sentiment analysis. Mainly, the huge volume of data, its variety and the speed in which the data is generated. Because of this, the need to use big data frameworks appeared in the case of sentiment analysis.

Several research proposed to enhance the efficiency of sentiment analysis extraction using big data frameworks. These research were conducted using machine learning techniques, lexicon based techniques or hybrid techniques.

The used machine learning techniques included Naïve Bayes, Logistic Regression algorithm and LMClassifier. Among these methods, the Naïve Bayes was the most used due to its high accuracy rates in comparison with the other methods [24].

In [3], the scalability of Naïve Bayes in handling large datasets was evaluated on a MapReduce framework. The authors used their own implementation of the algorithm, as shown in Figure 1, to help them in controlling the procedures of the analysis. The accuracy of the implemented code was nearly 80% with high scalability. The authors built their modules on the top of MapReduce framework. The modules were the work flow controller (WFC), the data parser, the user terminal and the result collector. The steps of the classification were as follow: first, the WFC module was utilized for performing the training job to build the model. Then, a combiner was used to merge the test data with the model, which produced intermediate results. Finally, the reviews were classified by calculating the probabilities of each review. The average of accuracy of the system was around 82%.

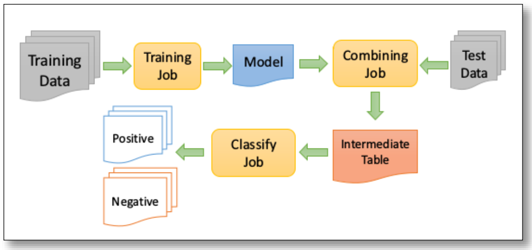

Figure 1: Naïve Bayes Evaluation system on Hadoop [3].

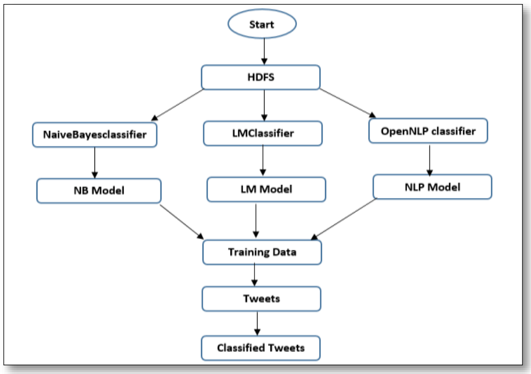

Figure 2: Classification using Naïve Bayes, LMC and OpenNLP classifiers [24].

The researchers in [24] utilized the Hadoop framework, specifically Apache Mahout, to create a model for clustering and classifying text into positive or negative classes. The classification of the text was based on extracted emotions from text using Naïve Bayes, OpenNLP and LMClassifier. The steps of the analysis are shown in Figure 2. It has been found that the accuracy of the NaiveBayes classifier outperformed LM Classifier and OpenNLP. However, for better execution time, the three classifiers were combined in the research to produce the results of the sentiment classification.

In addition, sentiment analysis was applied on a twitter data set, a movie data set, in [25] using the MapReduce framework. The method utilized the SentiWordNet, which is a trained dictionary that contains different words with their synonym and polarity. The focus of the research was to measure the effect of considering the emoticons in the preprocessing step on the sentiment analysis results. Each emoticon was replaced by a word that represent this emotion. The used preprocessing steps, shown in Figure 3, were removal of URL's, removal of special symbol, converting emotions, removal of username, removal of hashtag and removal of additional white spaces. The results indicated that preprocessing with considering the emoticons increased the accuracy. In addition, considering the emotions in the preprocessing reduced number of neutral results by classifying them correctly into positive and negative sentiments.

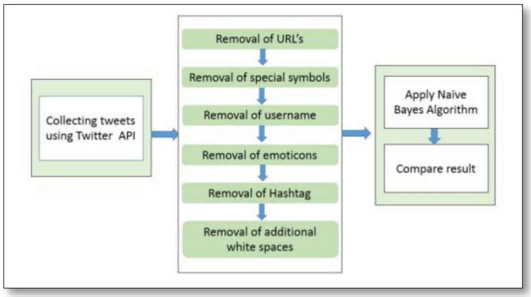
Collecting tweets using Twitter API
Removal of URL's
Removal of special symbols
Removal of username
Removal of emoticons
Removal of Hashtag
Removal of additional white spaces
Apply Naive Bayes Algorithm
Compare result

Figure 3: Sentiment analysis of Twitter Data-set using Naïve Bayes [25].

Other methods applied dictionary based methods for sentiment analysis. The used dictionaries included Positive context, Negative context, Positive word, Negative word, Prohibited words, sentiment-bearing words, lexicon of a sport event and AFINN dictionary with semantics using the WordNet. Additional analysis included syntactic analysis, Morpheme analysis and the Prohibited words analysis.

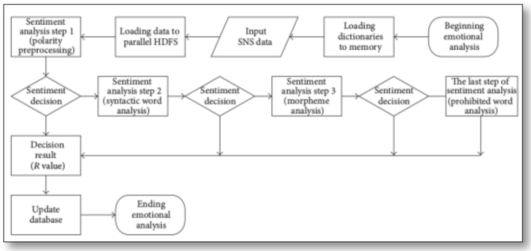
Sentiment analysis step 1 (polarity preprocessing)
Loading data to parallel HDFS
Input SNS data
Loading dictionaries to memory
Beginning emotional analysis
Sentiment decision
Sentiment analysis step 2 (syntactic word analysis)
Sentiment decision
Sentiment analysis step 3 (morpheme analysis)
Sentiment decision
The last step of sentiment analysis (prohibited word analysis)
Decision result (R value)
Update database
Ending emotional analysis

Figure 4: The emotion analysis using four emotion analysis functions [26].

In [26] authors proposed a parallel sentiment analysis method based on Hadoop framework. The method achieved very good results for extracting sentiments from unstructured texts using four functions for sentiment analysis. The functions, as shown in Figure 4, where the polarity preprocessing function, the Syntactic word analysis function, the Morpheme analysis function and the Prohibited words analysis function. The system achieved accuracy results that were close to manual analysis results. In addition, the time was linearly increasing as the data size increases, which indicates a high scalable system.

Moreover, the research utilized five types of dictionary for sentiment analysis. The first dictionary was the “Positive context” dictionary, which was used for polarity preprocessing and contained a set of positive context patterns for computing the positive contexts number in the sentence. The second dictionary was the “Negative context” dictionary, which was used for polarity preprocessing and contained a set of negative context patterns for computing the positive contexts number in the sentence. The third was the “Positive word” dictionary, which was used for syntactic and morpheme analysis. It contained a set of positive word patterns for computing the positive words number in the sentence. The fourth was the “Negative word” dictionary, which was used for syntactic and morpheme analysis. It contained a set of negative word patterns for computing the negative words number in the sentence. The final dictionary was the “Prohibited word” dictionary, which was used for analyzing Prohibited words. It contained a set of prohibited words for computing the prohibited words number in the sentence [26].

In [27], a dictionary based sentiment analysis technique was proposed based on the Hadoop MapReduce framework. The used dictionary contained sentiment-bearing words. The technique was evaluated using accuracy and execution time. The focus of the research was to develop a scalable technique to handle large datasets efficiently. Although machine-learning techniques provided more accurate results than dictionary-based techniques, they were not used since they require more time in training and model building.

The researchers in [28] proposed a system for analyzing sentiments at the entity level in the tweets. The research utilized the RDF graphs to analyze a tweet’s lexicon of a sport event to classify the tweets in either positive or negative.

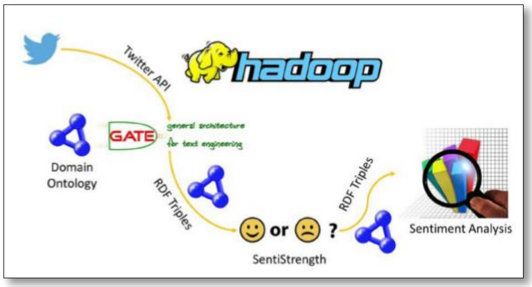

Figure 5: Elements of sentiment analysis system based on RDF graphs [28].

The research included three analyses: the first was the analysis of the relation between the number of tweets and the match scores during the championship. The second was the analysis of the relation of the tweet polarity and the score of the match. The last analysis studied the relation between tweets number and the sentiment value. The proposed method included four elements as shown in Figure 5. The first element, the domain ontology, described the context of the analysis by defining names and synonyms of entities, which will be recognized in the tweets. The second element, the tweet extraction and analysis algorithms, used to search for relevant terms in the tweets during the analysis process. The third element, the NLP analysis algorithm was developed by GATE, which is a Java tool for NLP analysis. The last element was the sentiment calculation algorithm, where the sentiment was calculated using SentiStrength according to the needed aggregation level [28].

The proposed approach in [11] utilized dictionary based technique for handling sentiment analysis on MapReduce framework. The approach analyzed the sentiment in the tweets by enriching the AFINN dictionary with semantics using the WordNet. The used semantic relation was the synonymy. The system, shown in Figure 6, included four main steps: the text preprocessing, the using of AFINN dictionary to determine words synonyms, assigning weights to words and finally the classification. Each word in the tweet was assigned a weight and a threshold on the sum of weights was used to determine the polarity of the tweet. The approach achieved good results of classification rate and error rate.

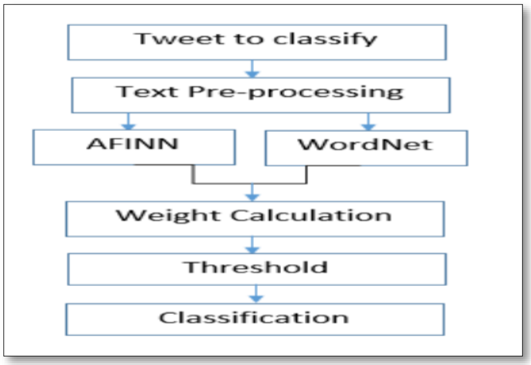

Figure 6: The steps of the semantic sentiment analysis method using WordNet and modified AFINN [11]

The authors in [4] proposed a technique to classify the tweets in two approaches, to classify it into positive, negative and neutral or to classify it into positive and negative. Before classification, the following preprocessing were applied: tokenization, removing stop words, removing URL removing numbers and negation handling. The first approach classified the tweets by computing the similarity of each tweet with three documents. Each document represented a class, and contained words that represent the classes. According to the similarity results, the tweets were assigned the class of the document, which it had the highest values of semantic similarity with it, Figure 6. The second approach, Figure 7, utilized a new formula to classify the tweets, which computed the semantic similarity between the tweets words and the two words "positive" and "negative" [4].

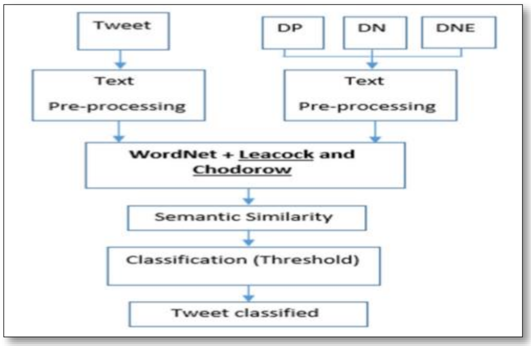

Figure 7: The classification steps using WordNet plus Leacock and Chodorow [4].

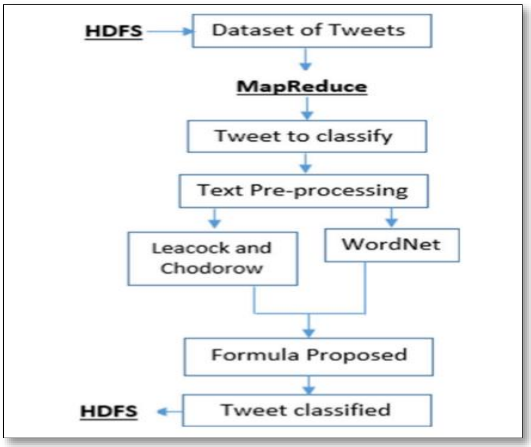

Figure 8: The second classification steps using WordNet plus Leacock and Chodorow [4].

Besides the machine learning and lexicon-based techniques, a hybrid technique was applied in [29], which combined the Logistic Regression algorithm with opinion lexicon.

The research in [29] proposed a system of two components, lexicon builder and a sentiment classifier, which was built for analyzing sentiment polarity on the MapReduce framework. The Logistic Regression algorithm was used as a classifier. The lexicon was built using Hadoop and HBase frameworks. The used lexicon was opinion lexicon, which constructed from a list of positive and negative words for twitter. The system, shown in Figure 9, achieved good scalability and accuracy results.

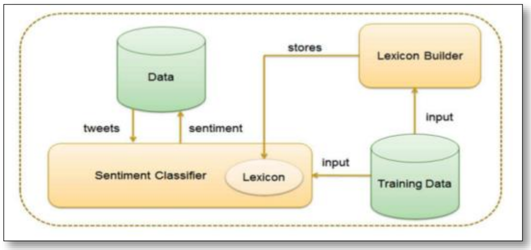

Figure 9: Architecture of the distributed sentiment analysis system [29].

The mentioned sentiment analysis techniques focused on utilizing the MapReduce framework for scalable data analysis. The description of used datasets, used algorithms, evaluation measures and results are summarized in Table 1. Most of these techniques utilized twitter data sets since it is the most used social network for sentiment analysis. The Naïve Bayes algorithm was the most used algorithm for sentiment analysis because of its high accuracy rate. However, lexicon based techniques were the interest of some researcheres because of its fast execution time in comparison with machine learning techniques. NLP methods were used to enhance the results of the analysis such as semantic analysis. Different preprocessing steps were applied to achieve higher accuracy rates such as tokenization, removing stop words, removing URL, removing special symbol, removing numbers, negation handling, converting emoticons and removing usernames.

5 Conclusion

The sentiment analysis is considered as a critical task in many applications. As the size of data increases, high-scalable sentiment analysis techniques are required to process the massive volume of data. Using MapReduce as a parallel environment provides scalable and efficient framework for sentiment analysis. Thus, several techniques utilized the MapReduce framework for extracting sentiments from large data sets. This paper reviewed the approaches and techniques of sentiment analysis that incorporated MapReduce framework. Most of these techniques were applied on Twitter data sets, in addition to reviews of Cornell University movie and Amazon movie review. The reviewed techniques achieved high scalability results. In addition, the accuracy was in high rates in the range 61%- 82%. The Naïve Bayes achieved the highest rate in accuracy; however, other machine learning techniques need to be investigated. Lexicon based algorithm achieved better results in terms of execution time.

The research of sentiment analysis on MapReduce still in need to speed up the analysis performance and increase the accuracy of the results.

REFERENCES

[1] T. White, Hadoop: The Definitive Guide, 4th Edition, California, USA: O'Reilly Media, 2015.

[2] B. Marr, "/big-data-," 30 9 2015. [Online]. Available: https://www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#4d9f09f717b1.

[3] B. Liu , E. Blasch, Y. Chen, D. Shen and G. Chen, "Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier," in *2013 IEEE International Conference on Big Data*, USA, 2013.

[4] Y. Madani, M. Erritali and J. Bengourram, "Sentiment analysis using semantic similarity and Hadoop MapReduce," *Knowledge and Information Systems*, pp. 1-24, 2018.

[5] H. Kaur, V. Mangat and N. Nidhi, "A Survey of Sentiment Analysis techniques," *International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, pp. 921-925, 2017.

[6] M. Edison and A. Aloysius , "Concepts and Methods of Sentiment Analysis on Big Data," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 5, no. 9, pp. 16288-16296, 2016.

[7] V. Patel, G. Prabhu and K. Bhowmick, "A Survey of Opinion Mining and Sentiment Analysis," *International Journal of Computer Applications*, vol. 131, no. 1, pp. 24-27, 2015.

[8] Y. MADANI, J. BENGOURRAM, M. ERRITALI, B. HSSINA and M. Birjali, "Adaptive e-learning using Genetic Algorithm and Sentiments Analysis in a Big Data System," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 8, no. 8, pp. 394-403, 2017.

[9] W. Medhat, A. Hassan and H. Korashy , "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, p. 1093–1113.

[10] M. Biltawi, W. Etaiwi, S. Tedmori, A. Hudaib and A. Awajan, "Sentiment Classification Techniques For Arabic Language: A Survey," in *2016 7th International Conference on Information and Communication Systems (ICICS)*, Irbid, Jordan, 2016.

[11] Y. Madani, E. Mohammed and B. Jamaa, "A Parallel Semantic Sentiment Analysis," in *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*, Rabat, 2017.

[12] O. Appel, F. Chiclana, J. Carter and H. Fujita, "A hybrid approach to the sentiment analysis problem at the sentence level," *Knowledge-Based Systems*, pp. 110-124, 2016.

[13] M. Biltawi, G. Al-Naymat and S. Tedmori, "Arabic Sentiment Classification: A Hybrid Approach," in *2017 International Conference on New Trends in Computing Sciences*, Amman, Jordan, 2017.

[14] M. Shiha and S. Ayvaz, "The effects of emoji in sentiment analysis," *International Journal of Computer and Electrical Engineering*, vol. 9, no. 1, pp. 360-369, 2017.

[15] M. Huq, A. Ali and A. Rahman, "Sentiment analysis on Twitter data using KNN and SVM," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, pp. 19-25, 2017.

[16] Z. Jianqiang and G. Xiaolin, "Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis," in *IEEE Access*, 2017.

[17] E. Haddia, X. Liu and Y. Shi, "The role of text pre-processing in sentiment analysis," in *Information Technology and Quantitative Management (ITQM2013)*, 2013.

[18] S. Hassan, F. Miriam, H. Yulan and A. Harith, "On stopwords, filtering and data sparsity for sentiment analysis of Twitter," *LREC 2014, Ninth International Conference on Language Resources and Evaluation*, p. 810–817, 2014.

[19] T. Singh and M. Kumari, "Role of Text Pre-processing in Twitter Sentiment Analysis," *Procedia Computer Science*, pp. 549-554, 2016.

[20] W. Sharif, N. Samsudin, M. Deris and R. Nase, "Effect of Negation in Sentiment Analysis," in *2016 Sixth International Conference on Innovative Computing Technology (INTECH)*, Dublin, Ireland, 2016.

[21] H. Saif, Y. He, M. Fernandez and H. Alani, "Semantic Patterns for Sentiment Analysis of Twitter," in *International Semantic Web Conference*, 2014.

[22] S. Tartir and I. Abdul-Nabi, "Semantic Sentiment Analysis in Arabic Social Media," in *Journal of King Saud University – Computer and Information Sciences*, 2017.

[23] M. Biltawi, W. Etaiwi, S. Tedmori and A. Shaout, "Fuzzy based Sentiment Classification in the Arabic Language," in *Intelligent Systems Conference 2018*, London, UK, 2018.

[24] V. Chauhan and A. Shukla, "Sentimental Analysis of Social Networks using MapReduce and Big Data Technologies," *IJCSN International Journal of Computer Science and Network*, vol. 6, no. 2, pp. 120-130, 2017.

[25] H. Parveen and S. Pandey, "Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, Bangalore, India, 2016.

[26] I. Ha, B. Back and B. Ahn, "MapReduce Functions to Analyze Sentiment Information from Social Big Data," *International Journal of Distributed Sensor Networks*, pp. 1-11, 2015.

[27] R. Ramesh, G. Divya , D. Divya, M. Kurian and V. Vishnuprabha, "Big Data Sentiment Analysis using Hadoop," *IJIRST –International Journal for Innovative Research in Science & Technology*, vol. 1, no. 1, pp. 92-96, 2015.

[28] C. González, J. García-Nieto, I. Navas-Delgado and J. Aldana-Monte, "A Fine Grain Sentiment Analysis with Semantics in Tweets," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 3, no. 6, pp. 22-28, 2016.

[29] V. N. Khuc, C. Shivade, R. Ramnath and J. Ramanathan, "Towards building large-scale distributed systems for Twitter," *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC ’12)*, p. 459–464, March 2012.

Table 1: Comparison of Sentiment Analysis Techniques based on MapReduce.

Ref	Year	Data set used	Polarity used	Sentiment analysis method	Evaluation measures	Results
[3]	2013	Two data sets were used, the Cornell University movie review, which had 1000 positive reviews and 1000 negative reviews. The second data set was the Amazon movie review dataset, which structured in eight lines for each review. This data set contained five points rating system, which was converted into two, positive and negative using 3.0 as a threshold.	Positive/negative	Naïve Bayes	Accuracy Computation time Throughput of the system	The results indicated increase of the system throughput. The average of accuracy was below 82%. [4]
[24]	2017	The used datasets are the Sanders Twitter Dataset, which contained 1400 tweets. The second data set was the Stanford Twitter Sentiment, Test Set (STS-Test).	Positive/negative	Naïve Bayes, OpenNLP and LMClassifier	Accuracy	61 to 72.83 for openNLP 62 to 69.79.83 for LMClassifier 64.04 to 75.66 for NaiveBayes
[25]	2016	A Twitter Dataset	Positive/negative/neutral	Naïve Bayes	NA	The results show that applying processing with considering emoticons produce results that are more accurate. Specifically, number of neutral sentiments were reduced.
[26]	2015	Five Twitter data sets that was gathered by the “Topsy” application programming interface (API).	Positive/negative/neutral	Naïve Bayes, OpenNLP and LMClassifier	Execution time Accuracy	Accuracy, 70% for positive sentiments 15.8% for negative sentiments 14.2% for neutral sentiments
[27]	2015	The data sets was a collection of tweets, gathered by the Flume application.	Positive/negative/neutral	dictionary based technique	Precision Recall F-measure Accuracy Execution time	Precision, 66.666% Recall, 100% F-measure, 79.95% Accuracy, 75%
[28]	2016	Analyzed the sentiments in tweets related to specific sport event (the Phillips 66 Big 12 Men’s Basketball Championship of 2014).	Positive/negative	ontology-based NLP process	The Pearson product-moment correlation coefficient	
[11]	2017	Tweets data set was collected using the JAVA Twiter API (Twitter4j) and the Apache Flume.	Positive/negative/neutral	Dictionary-based approach (the Wordnet plus the AFINN dictionary)	classification and error rate of the classification of the tweets	Without text pre – processing Classification Rate, 0.74 Error Rate, 0.26 With text pre – processing Classification Rate, 0.82 Error Rate, 0.18
[4]	2018	Two data sets were used. The first contained tweets, which was collected using Twitter4j. The second dataset was collected using Flume.	Positive/negative/neutral & Positive/negative	Leacock and Chodorow and WordNet	Accuracy, the error rate, precision, accuracy and F1 measure	**First approach** Accuracy: 91% Precision: 8% Recall: 85% F1-score: 82% **Second approach** Accuracy: 93% Precision: 81% Recall: 84% F1-score: 83%
[29]	2012	The data set consist of 384397 tweets. 232442 tweets had smileys and 151955 tweets had frownies. The produced lexicon had 2411 positive words/phrases, and 1018 negative words/phrases.	Positive/negative/neutral	Lexicon-based and Adaptive Logistics Regression	Accuracy Execution time	Baseline, 55.4% Lexicon-based, 72.1% Lexicon-and-learning-based, 73.7%