COMPUTER SCIENCE | RESEARCH ARTICLE

Aspect-based classification of product reviews using Hadoop framework

Anisha P. Rodrigues¹, Niranjan N. Chiplunkar¹ and Roshan Fernandes¹

Abstract: The advancement of e-commerce along with the quick development of product review discussion in the most recent decade, an enormous measure of sentiment data or reviews are produced which made it practically difficult for a customer to take an educated buy choice. A good number of customers share their conclusions about the product during these days. These reviews have an important role in customers’ purchase-decision process. For a popular product there may be hundreds or thousands of reviews. This is difficult for a potential customer to go to each of them to make an educated choice on whether to buy the product or not. Also, this is difficult for manufacturer of the product to follow-up and to manage client opinions. In this scenario, the aspects-based sentimental analysis helps in analyzing the reviews and categorizing them into appropriate aspects. Aspect or feature refers to attributes or qualities of a product. The proposed work begins with collecting reviews from online shopping websites, identifying aspects and classifying opinion orientation of aspects with different sentiment analysis techniques using Hadoop framework. This paper proposes a new pattern-based method for aspect extraction and sentiment analysis which gives an accuracy in the

ABOUT THE AUTHORS

Dr. Anisha P Rodrigues Working as Assistant Professor in the Department of Computer Science and Engineering at NMAM Institute of Technology, Nitte, India. Her research interests include Natural Language Processing, Machine Learning, Big Data Analytics and Image Processing. She is a lifetime member of Indian Society for Technical Education.

Dr. Niranjan N Chiplunkar Working as Principal and Professor in the Department of Computer Science and Engineering at NMAM Institute of Technology, Nitte, India. He is a Fellow of Institution of Engineers (India) and member of several other Professional bodies like IEEE, CSI, and ISTE. His research interests include Data Mining, CAD for VLSI, Big Data Analytics and Natural Language Processing.

Dr. Roshan Fernandes Working as Associate Professor in the Department of Computer Science and Engineering at NMAM Institute of Technology, Nitte, India. His research interests include Mobile Web Services, Natural Language Processing and Machine Learning. He is a lifetime member of Indian Society for Technical Education.

PUBLIC INTEREST STATEMENT

With the growing availability and popularity of opinion rich sources such as online reviews, choosing the right product from huge product brands is difficult for the user. To improve the marketing and enhance the customer satisfaction, various retailer portals provide opportunity for the user to share their reviews on products and its aspects. It is difficult for the new customer and product manufacturer to understand interest of a customer about the product. In this situation, sentiment analysis helps the people to analyze the reviews and conclude whether it is good or bad. The aspect level sentiment classification helps new customers to know more interesting aspects of products. Aspect or feature refers to attributes of a product. This article presents new big data approach for aspect extraction and sentiment analysis of product reviews. It proposes a method which identifies aspects using pattern-based approach and classifies opinion orientation of the aspect. Entire work is carried out on Hadoop framework and the proposed pattern-based method gives better actuary.
range of 72 – 75%. The proposed work is implemented on Hadoop MapReduce framework and the results show that Hadoop Multi-Node cluster set up performs aspect level sentiment analysis in a shorter time compared to traditional techniques.

**Subjects:** Computer Engineering; Computational Linguistic & Language Recognition; Machine Learning – Design; Computing & IT Security; Computer Science (General)

**Keywords:** sentiment analysis; big data; Hadoop; support vector machine; Naïve Bayes classifier

1. Introduction

The web-based shopping has gained a lot of prevalence in recent years. This development is the consequence of different benefits on the web shopping which has over the traditional buying exercise. To mention few benefits—it saves customers’ time by allowing them to purchase a product with few mouse clicks, provides the flexibility to compare the prices and make a buying decision, availability of online stores. As the popularity of online shopping has increased, the products purchase and sale on the web are also increased. The problem of web shopping is that, it might be hard to get assistance from the experts to purchase an item. With the product details, it is difficult for the customer to make a purchase decision. To overcome this problem, the suppliers have enabled the blogs, review sites, forums through which the customer can express their opinion about the purchased item. These opinions or reviews are more important than the descriptions provided by the manufacturer because customers express their views about the purchased item in detail, what the product is about, how it functions, etc. The popular or trendy items usually get many reviews. Besides, numerous reviews are long and have just a couple of sentences containing sentiments on the item. This makes difficult for other customers to go through all reviews and make a purchase decision. This has encouraged research on sentiment analysis. Since most of the extracted reviews are in unstructured or semi-structured format, and having larger size, traditional data processing techniques fail in analyzing them. As the data size increases, analysis of the data becomes challenging. Hence there is a need of big data processing technique to analyze this huge data in an effective and efficient way. In recent times, the effective tool that has turned out to be productive in handling large datasets is Hadoop (White, 2012), which is thought to be proficient for storage and processing of large sets of data. So, the main idea behind this work is to define a system which uses distributed processing technique to extract knowledge from big data sentiments.

Sentiment or opinion is an idea, view, or state of mind, particularly one construct principally in light of feeling rather than reason. Sentiment analysis also called as opinion mining contemplates individuals’ estimations towards specific elements. Sentiment analysis is one of the significant errands of NLP (Natural Language Processing). Sentiment analysis can be performed at three levels: Document level, Sentence level, and Aspect level (Medhat et al., 2014). At Document level, entire document is treated as a single element and sentiment analysis is performed to check whether the document conveys positive or negative view. At sentence level, opinion summarization is performed at the sentence level. First, the sentence is classified into subjective or objective sentence. Second, the polarity of subjective sentence is classified. An objective sentence expresses factual data and subjective sentence expresses subjective observations and views. People use e-commerce sites to buy products online and then they post reviews or opinions about the products. In these sites, customers express their opinions on products what they buy and share their opinions on the product aspects/features. Such reviews can also be studied to understand the best and worst features of the product. This is called aspect level sentiment analysis. Aspect level classification involves several subproblems, namely, detecting appropriate entities, mining their aspects/features, and identifying if the review specifies positive or negative opinion on each aspect. Complete and detailed information
on entity and its aspects cannot be provided by document level or sentence level text classification procedure. This information is required for many business applications to make a better decision and to promote their business. So, the objective of this study is to design and develop an efficient algorithm to analyze real-time product reviews from retailer sites and perform aspect-based classification using Hadoop MapReduce framework.

Figure 1 shows the taxonomy of the sentiment analysis techniques as discussed in article (Medhat et al., 2014). Sentiment Classification techniques can be grouped into two types: Lexicon-based approach and Machine Learning approach. The Lexicon-based approach is based on sentiment words that are normally utilized as a part of communicating positive or negative opinion. It depends on a sentiment vocabulary which consists of known and precompiled sentiment words. This sort of methodologies utilizes SentiWordNet, which is a tagged dictionary with precalculated scores assigned for each term.

Machine learning approach uses machine learning algorithms in order to perform sentiment analysis. Machine learning approach further categorized into two types: Supervised and Unsupervised learning techniques. Supervised learning method depends on the labeled dataset which is used to training classification model and the classification is performed using machine learning algorithms. Partial and hierarchical clustering are the most commonly used unsupervised algorithms. In machine learning approach, accuracy of the sentiment analysis mainly depends on training dataset, from which the classifier is built. As the review data set size increases, the accuracy decreases if we use same training dataset. Since the Lexicon method does not rely on training data set, it is best suited for sentiment analysis. We have proposed a new pattern-based method for aspect extraction and opinion phrase extraction. Sentiment analysis is carried out using Lexicon-based method. The various steps followed while performing aspect level sentiment analysis include—extracting the reviews, preprocessing, POS tagging, aspect selection, and classifying reviews into positive or negative.

Section 2 provides the brief literature survey on different techniques used to perform sentiment analysis and aspect level sentiment analysis. In section 3, the preliminaries of the work are discussed. Section 4 provides the details of proposed architecture. The experimental results are discussed in section 5. Section 6 discusses about conclusion and future work.

2. Literature survey

2.1. Sentiment analysis

Increasing popularity of e-commerce has given chance to sentiment analysis. Sentiment analysis identifies and categorizes the opinion of a text or a review. Walaa Medhat et al. (2014) have performed a study on different techniques of sentiment classification. According to the survey, the two sentiment classification techniques are Lexicon-based approach and Machine Learning Approach. The machine learning approach further divided into supervised and unsupervised learning
approach. Fang & Zhan (2015) have performed sentiment analysis on product reviews using supervised machine learning algorithms at sentence level. The proposed method collected product reviews of different domains from Amazon website. Out of three classification methods, i.e. Naïve Bayes (NB), Support Vector Machine (SVM), and Random forest, random forest methods provide the good result on sentiment analysis. Moroes et al. (2013) have performed sentiment analysis at document level and compared the performance of SVM and artificial neural network (ANN). The datasets considered are from different domains, i.e. movie reviews, book reviews, GPS data, and camera reviews. The result concluded that ANN outperformed SVM in sentiment classification, but the training time of ANN exceeded the training time of SVM. Zhang et al. (2014) performed opinion mining on reviews obtained from the mobile users. The authors compared two classification methods namely SVM liblinear and NB multinomial. From the result it has been proved that NB multinomial performed better when compared to SVM liblinear. Hu & Liu (2004) and Potdar et al. (2016) have followed Lexicon-based approach to produce the summarization of product reviews collected from the Amazon website. Deshmukh & Tripathy (2018) have proposed a semi-supervised approach for sentiment analysis using modified maximum entropy method. The work concentrated on collecting sentiment words from one domain which can be used to predict the sentiment words of another domain. The study also concluded that the proposed method is best when compared to SVM and NB. Yang et al. (2012) utilized SVM method to classify the text into positive or negative. Kumar et al. (2016) performed a study on various classification methods like SentiWordNet, Logistic regression and NB approach. The study concluded that NB method outperforms the other two approaches. Chifu et al. (2015), proposed an unsupervised approach for aspect level sentiment analysis on product reviews. The proposed method used ant clustering algorithm to select the aspects of particular product. To classify the reviews, self-organizing map approach is proposed. Esparza et al. (2017) performed sentiment analysis on product reviews using various SVM kernels such as SVM linear, SVM radial, and SVM poly. Among the three, the linear kernel has achieved the greater accuracy. Kang & Park (2014) performed sentiment analysis to measure the customers’ opinion on mobile service using dictionary approach. The sentiment dictionary used to identify the keyword vector in the review sentence and the summarization of customer opinion is performed using multicriteria decision-making approach. Khan et al. (2017) proposed a semi-supervised method to perform opinion mining. The method combined Lexicon-based approach along with machine learning technique, i.e. support vector machine to achieve good results. Jinturkar & Gotmare presented a framework to perform sentiment analysis of review data using Hadoop framework. Assigning the values to the sentiment words are performed at mapper. Classification of polarity using Naïve Bayes is carried out at reducer. Maddikunta et al. (2020) have discussed the effect of dimensionality reduction techniques, namely, Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) on machine learning algorithms. Authors have concluded that classifier with PCA performs better compared to LDA. Gadekallu et al. (2019) proposed a location-based business recommendation system. Recommendation system’s quality is measured in terms of customer business ratio and average service time on bench mark datasets. Authors have proved that, quality of proposed recommendation system in benchmark datasets gives better results compared to existing K-nearest neighbor algorithm.

2.2. Aspect level sentiment analysis

Hu & Liu (2004) have considered the words tagged with noun as aspects and applied compactness pruning and redundancy pruning technique to filter out the words which are not aspects. Bafna & Toshniwal (2013) have proposed a method for selecting domain specific aspects. The method involved probabilistic approach and association rule mining. An aspect selection method proposed by Maharani et al. (2015) follows a syntactic pattern approach. It follows various patterns to extract single word, double word and tri word aspects. The proposed work concentrated only on extracting explicit aspects. Devasia & Sheik (2016) identified aspects with the help of dependency between the words of review sentences. Stanford dependency parser is used to identify the relationship between the words. To filter the selected aspect, stemming is applied so that aspects in plural form can be converted into singular form. Ghosh & Sanyal (2017) performed aspect level sentiment analysis selection using NB and SVM. The proposed method considered frequency of aspect and static measure called Term Frequency Inverse Document Frequency. Parlar & Özş (2016) proposed a new
aspect selection technique which depends on query expansion term weighting methods. The performance of proposed method is compared with chi-square method and document frequency difference techniques out of which the proposed method gained higher accuracy for Turkish reviews. Sharma & Dey (2012) carried out a comparative study on five different aspect selection methods. The result concluded that out of five methods, Gain Ratio is efficient as it gained highest accuracy. Chauhan & Singh (2017), proposed novel approach for aspect selection and used different filtering methods, namely, distance-based filtering, subset filtering, and superset filtering to filter redundant and basic features. Sentiment analysis is performed using Lexicon-based method to find the sentiment word associated with extracted aspects. Li et al. (2015) presented a model for Aspect-based sentiment analysis which follows clustering model to identify the aspects. The aspect group is formed by identifying their synonyms. The work concentrated on finding opinion of dynamic sentiment ambiguous adjectives with the help of dependency pattern rules. Htay & Lynn (2013), proposed novel approach for finding aspects and extraction opinion words of product reviews. Authors have proposed patterns for extracting aspects and opinion phrases and carried out experiments on five electronic product reviews extracted from amazon web site. Proposed pattern showed slightly better results compared to Hu & Liu (2004) approach. Chifu et al. (2015) proposed an unsupervised method for aspect level opinion mining on product reviews. The proposed method used ant clustering algorithm to select the aspects of particular product. To classify the reviews, self-organizing map approach is proposed. Gadekallu et al. [31] have discussed domain specific Aspect level sentiment analysis on Movie review dataset. Authors concluded that integrating syntactic information to the model will give better results in sentiment analysis.

In the survey of opinion mining, Guo et al. (2014) suggested that, sentiment analysis algorithms can be improved further by applying big data processing techniques to analyze larger volume of data. Table 1 shows the literature survey summary of aspect level sentiment classification. The study and literature survey shows that the sentiment analysis is performed on smaller data sets at various levels using different classification algorithms. The datasets considered for the experimental purpose are static review sets and smaller in size. As the dataset size increases, the algorithms used for sentiment analysis may not scale up well. So, there is a need of effective framework to process large amount of data. Hadoop framework is the preferred computing framework for processing larger datasets since Hadoop clusters are highly available and operates in a fault tolerant manner. To gain better accuracy and to process data in a distributed manner, we have proposed pattern-based method for aspects and opinion phrase extraction.

3. Preliminaries

3.1. Hadoop framework

Hadoop is openly accessible java-based software, initially implemented by Doug Cutting and Mike Cafarella in 2005. After the advancement, it has become an enlisted trademark of the Apache Software Foundation. The advantage of Hadoop is that, it provides an intense distributed platform for processing and managing Big Data. It executes applications on massive clusters of commodity hardware. Using Hadoop, it is possible to process many terabytes of data on large number of nodes. The significant favorable position of Hadoop system is that it gives reliability and high availability. Hadoop has two major components: Hadoop Distributed File System (HDFS) for storing purpose and MapReduce for processing purpose (White, 2012).

MapReduce, a programming model, is mainly utilized for analyzing and producing expansive datasets with the help of parallel and distributed algorithm on a cluster. For the most part it follows three steps: dividing, processing, and merging. MapReduce works in two stages. They are called map phase and reduce phase. Both the phases have key-value pair as input and output. The map phase takes the input, processes it and produces an intermediate key-value pair which will be given as input to the reduce phase. The job of the reduce phase is to merge every intermediate value related with the same key. The MapReduce structure comprises of two processes that are JobTracker and TaskTracker. It is the responsibility of JobTracker to handle users’ request. The
| Author                  | Techniques used                  | Key findings                                                                 | Limitations                                                                                                                                 |
|------------------------|----------------------------------|------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Hu & Liu (2004)        | Lexicon-based method             | Proposed new approach for feature-based summarization of product reviews. This system performs tasks namely tagging, opinion word extraction, identifying opinion orientation of opinion words, and predicting the orientation of opinion sentences and summary generation. | Algorithm lagged in finding out the strength of opinions, exploring opinions having nouns, verbs, and adverbs and Pronoun resolution. |
| Htay & Lynn (2013)     | Lexicon-based method             | Proposed approach for finding aspects and extracting opinion words of product reviews. Authors have proposed patterns for extracting aspects and opinion phrases. Proposed pattern showed slightly better results compared to Hu & Liu (2004) approach. | Extraction of implicit features is not addressed.                                                                                           |
| Bafna & Toshniwal (2013)| Lexicon-based method             | Proposed dynamic domain specific aspect-based summarization of customers' opinions on online products. | Algorithm lagged in learning opinions expressed with nouns, verbs, and adverbs. The proposed approach for aspect extraction and opinion polarity detection needs to be improved. |
| Maharani et al. (2015) | Syntactic pattern-based method   | Proposed syntactic pattern for aspect extraction from product reviews and tested with different pattern combination. | Proposed approach did not succeed in the extraction of both implicit and explicit aspects. The experimental results showed that the syntactic pattern approach had several weaknesses that need to be improved. |
| Devasia & Sheik (2016) | Semantic-based approach          | Proposed semantic-based approach for extracting product aspects. Recursive deep analyzer is used to find out opinion orientation of Product review sentences. | Sentiment analysis is done by using recursive deep analyzer which needs to be further improved by applying various machine learning techniques. Proposed method fails in capturing implicit aspects and other contextual information used for classification. |
| Ghosh & Sanyal (2017)  | Naive Bayes and Support Vector Machine | Authors have done comparative study on sentiment analysis of product reviews using Naive Bayes and Support Vector Machine and proved that support vector machine classifier performs better than Naive Bayes classifier for various mixture of features. | Authors have done only a preliminary study on sentiment analysis. More feature selection techniques need to be adopted to improve the classifier accuracy. |
| Parlar & Özel (2016)   | Naive Bayes Multinomial classifier | Proposed feature selector method which gave accuracy better than Document Frequency Difference and Chi-Square methods. Sentiment analysis Turkish product review is done using Naive Bayes Multinomial classifier. | Authors have not considered reviews written in different languages.                                                                                                                                 |

(Continued)
| Author          | Techniques used | Key findings                                                                                                                                                                                                                                                                                                                                 | Limitations                                                                                                                                                                                                                           |
|-----------------|-----------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Sharma & Dey (2012) | Survey paper    | Authors have done comparative study on different feature sector methods and sentiment analysis for online movie review dataset. Feature sector comparison done using five methods namely Document Frequency, Information gain, Gain Ratio, CHI statistics and Relief-F and seven machine learning-based classification techniques Naïve Bayes, Support Vector Machine, Maximum Entropy, Decision Tree, K-Nearest Neighbor, Winnow, Adaboost. | The experimental results presented in this paper show that Gain Ratio gives the best performance for sentimental feature selection, and SVM performs better than other techniques for sentiment-based classification.                          |
| Chauhan & Singh (2017) | Lexicon-based method | Proposed frequent item set mining approach to extract candidate aspects and used different filtering methods namely, distance-based filtering, subset filtering and superset filtering to filter redundant and nonaspect phrases.                                                                                     | More filtering techniques need to be incorporated to improve aspect extraction accuracy.                                                                                                                                               |
| Li et al. (2015)     | Lexicon-based method | Proposed Sentiment Summarization on Product Aspects (SSPA) approach which has incorporates different aspect sentiment analysis tasks namely, Product feature extraction, grouping features into meaningful aspects, Sentiment orientation, and finding out sentiment strengths. Experiments proves that each module of SSPA performs well. | More types of features including verbs, adjectives and implicit features need to be addressed.                                                                                                                                       |
| Chifu et al. (2015)  | Unsupervised method | Proposed ant clustering algorithm to select the aspects of particular product. To classify the reviews, self-organizing map approach is proposed.                                                                                                                                                                                                 | As further work, we will experiment our method on other opinion mining data sets, for instance, the movie review data set. Algorithm failed in addressing opinion reviews with multiple opinions having opposite polarity. Other domain datasets need to be experimented. |
JobTracker will coordinate the activity by assigning these tasks to execute on various datanodes. The TaskTracker which is residing on each DataNodes will manage the execution of each subtask. It also sends the progress report to the Job Tracker.

Hadoop Distributed File System was created utilizing distributed file system outline. The prominent features associated with HDFS are scalable, distributed and fault tolerant. HDFS maintains a master/slave architecture additionally it has the idea of a block. The master/slave pattern comprises of single NameNode and several DataNodes. NameNode deals with the filesystem and it contains greater part of its metadata in RAM. It has the information about all the DataNodes on which the blocks of a given file are stored and also handles the file system operations like creating, renaming, and closing directories and files. Apart from NameNode and DataNode, there is a third process called Secondary NameNode which works simultaneously with NameNode as a helper process. It periodically reads the information regarding file system and metadata from the NameNode’s RAM and composes it into the hard disk or the file system. Figure 2 depicts the Hadoop framework.

3.2. Hadoop multinode cluster

A Hadoop cluster can be thought of as a gathering of freelance systems associated via a dedicated network to function as a centralized information processing asset. The computational computer cluster is the one which disseminates workload of data analysis process over different nodes of cluster which work altogether for processing data in distributed manner. In Hadoop cluster framework, aside from the network nothing is shared between the cluster nodes. This shared nothing paradigm of cluster decreases the latency of processing, hence at the point when there is a need to analyze questions on immense measure of data the processing time over cluster is totally reduced. Hadoop Multi-Node cluster follows master/slave architecture. Master node stores the data in HDFS and executes parallel calculation on data utilizing MapReduce. There are three nodes in Master node: Job Tracker, NameNode and Secondary NameNode (Subramaniyaswamy et al., 2015). It is the job of JobTracker to control the distributed processing of enormous data with the help of MapReduce. NameNode monitors all the information of the documents stored on HDFS, i.e. the metadata. It stores the information like accessing time of the document, distribution of documents over the Hadoop cluster and details about the user accessing a document on current time. The secondary NameNode contains the NameNode data backup. Slave node part of Hadoop cluster is in charge of storing and processing the data. Both DataNode and TaskTracker service run on each slave node in order to communicate to the master node. Figure 3 depicts the Hadoop cluster architecture.

Figure 2. Hadoop framework.
4. Proposed work
The main aim of proposed framework is to process the customer reviews from e-commerce web site. The customers will read the reviews before finalizing the purchase decision. But it is not really conceivable to go through all the reviews. Hence there is a demand for sentiment classification. In our study, the real-time reviews are extracted using Jsoup parser and proposed new tri-gram pattern for explicit aspect extraction and bi-gram pattern is used for finding out opinion phrases. Entire work is implemented using two MapReduce stages. Aspect extraction is performed in the first MapReduce stage, whereas identifying opinion phrases and sentiment classification is performed in the second MapReduce stage. Figure 4 depicts the proposed architecture for Aspect level Sentiment Analysis. The proposed approach is explained below:

4.1. Review extraction
In the proposed work the reviews of electronic gadgets are extracted from e-commerce web sites using Jsoup. Jsoup is a library also known as Java HTML parser for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jQuery-like methods. Jsoup is designed to deal with all varieties of HTML found in the wild; from pristine and validating, to invalid tag-soup; Jsoup will create a sensible parse tree. For Jsoup, one single seed URL of product web page is given as input which will then explore through all the web pages to obtain the customer reviews. The extracted reviews are stored in HDFS which will be given as input to the aspect level sentiment analysis process. Figure 5 shows the flow of review extraction.

4.2. Pre-processing of raw data
The preprocessing step involves appropriate fragmentation and cleaning of data which is performed using short/slang word replacement, multiple characters removal, and dealing with emoji’s.

The review text may contain short word opinions such as nic, gud, g8, and many more are converted into full forms. For short words and slang word replacement, the words in the review text are compared with slang word corpus. If the word in the review text matches with the slang word in the corpus then it is replaced with appropriate, meaningful word. It also replaces the elongated words such as “suppeerrrb” with “super”. The review text is also examined to eliminate the special characters like “+,” “-,” “=,” “#,” “$,” “*,” etc. An emoticon is a short form of emotion icon, which is the visual portrayal of facial expression, used to communicate the opinion, attitude, or emotion. The emoticons which express the feelings like happy, sad, angry, thumbs up, and thumbs down are considered. Such emoticons are identified and converted into text using Java emoji library.

4.3. Parts of speech (POS) tagging
POS tagging is utilized to mark each word in the review sentence with its legitimate part of speech, i.e. it labels the words as verb, adverb, conjunction, preposition, noun, adjective, and interjection. Aspects
of a product are usually nouns and opinions are usually adjectives, adverbs or verbs. In this work Stanford POS tagger (Reddy et al., 2020) is used to label all the online review sentences. Table 2 shows the Stanford POS tags and corresponding description. For example, the sentence given to before POS tagging: “This phone has good battery life and speaker voice is good.” After POS tagging: This/DT phone/NN has/VBZ good/JJ battery/NN life/NN and/CC Speaker/NN voice/NN is/VBZ good/JJ.

4.4. Aspect extraction

The main focus of the proposed work is on performing aspect extraction and sentiment analysis using pattern-based method. Pattern analysis is performed to extract explicit aspects syntactic pattern that corresponds to the aspects of product. For the experiment purpose, we have taken annotated review datasets from Hu and Liu research (Hu & Liu, 2004) which has reviews on electronic products such as digital camera, cellular phone, mp3 player, and DVD. These reviews are extracted from e-commerce web site, namely, amazon.com. Also, we included the
prepossessed real-time product reviews extracted from amazon.com. Some of the example datasets are as follows:

(i) Camera [+2] [u]##this camera is perfect for an enthusiastic amateur photographer.
(ii) picture [+3] [u], macro[u] [+3] ##the pictures are razor-sharp, even in macro.
(iii) size [+2] [p]# it is small enough to fit easily in a coat pocket or purse.
(iv) weight [+1] [p]# it is light enough to carry around all day without bother.

Following are the descriptions of symbols used in an annotated review dataset.

xxxx[+|-n]: xxxx is a product feature.

[+]n: Positive opinion, n is the opinion strength: 3 strongest, and 1 weakest.

[-n]: Negative opinion.

##: start of each sentence and each line is a sentence.

[u]: feature not appeared in the sentence that is explicit feature.

[p]: feature not appeared in the sentence that is implicit feature.

In a sentence “camera [+2] ## this camera is perfect for an enthusiastic amateur photographer”, camera is explicit feature with positive opinion strength 2.

In e-commerce sites, customers express their reviews about products by highlighting some of the aspects of the products. For instance, in the review sentence “This phone has good battery life” customer has shared review on product aspect “battery life” and the opinion word is “good.” In our work, we, first identify product feature from product review. Many of the product aspects are nouns and noun phrases. Author Htay & Lynn (2013) proposed tri-gram pattern for extracting opinionated phrases and aspects. Table 4 shows the pattern used by Htay for aspect-based sentiment analysis. Here DT represents determiner, JJ represents adjective, RB represents adverb, VB represents verb, NN represents noun. This pattern extracts both opinion words and aspects. First four patterns are for extracting aspects which come with opinionated words and last four patterns are for opinion words. These trigram patterns extract only the noun aspects which comes with opinion words like adjectives, adverbs and verbs. But there are cases where aspects come other than opinion words. For instance, in the POS tagged sentence, “sometimes (RB) battery (NN) can (VB) get (VB) very (RB) hot (VB)” aspect “battery” comes with RB (adverb) and VB (verb). Htay pattern will fail to extract these kinds of aspects. So, to extract explicit aspects, we have proposed tri-gram syntactic pattern. Table 3 shows the proposed pattern for explicit aspect extraction. The aspects are derived from current POS tag with their (current-1) POS tag and (current+1) POS tag. Out of the extracted features, several nonfeatures phrases are also extracted. So, to eliminate these non-feature phrases, we have used Apriori algorithm to find frequent candidate features. In the Apriori algorithm, each and every extracted feature is taken as one transaction and sentences that contain those features are taken as item sets. Minimum support is calculated as a ratio of sentences contains all the features to total number of extracted features. If Sentence count of each feature is greater than minimum support, then such features are taken as candidate features. Algorithm 1.1 briefs on proposed aspect extraction. In the algorithm, Feature extraction as per the proposed pattern is done in Map phase and nonfeature filtering is done in Reduce phase.

**Algorithm 1.1: Aspect extraction**

**Data:** dataset: = annotated review datasets

**Results:** Aspects and its related Sentences stored in key-value pair form

Keywords used: fw = first Word, sw = Second Word, tw = Third Word

**function** Map (docID, line)

```plaintext```
for all line ∈ D do
```

Page 12 of 22
Table 3. Proposed pattern for aspect extraction

| First Word | Second Word | Third Word | Example |
|------------|-------------|------------|---------|
| NN/NNS     | -           | -          | Its/PRP having/VB great/JJ display/NN. |
| NN/NNS     | NN/NNS      | NN/NNS     | The DT nice/JJ thing/NN is/VB that/IN it/PRP uses/VB the_DT SD/NN memory/NN card/NN. |
| JJ         | NN/NNS      | NN/NNS     | Excellent/JJ picture/NN quality/NN and/CC so/RB simple/JJ to use/VB. |
| JJ         | NN/NNS      | RB/VB/VNZ  | The DT battery/NN lasts/VBZ little/RB. |
| DT         | NN/NNS      | -          | This DT camera/NN is/VBZ so/RB light/JJ and/CC tiny/JJ that/IN I/PRP literally/RB carry/VB it/PRP everywhere/RB. |
| DT         | NN/NNS      | IN/NN/NNS/DT | I like the DT lens/NN of /IN this camera. The DT pictures/NNNS some/DT time/NN turn/VB out blurry/JJ. |
| DT         | NN/NNS      | VB/VBZ     | The /DT player /NN keeps/ VB giving/VB random/JJ errors /NN. |
| RB         | NN/NNS      | VB/VBZ     | Sometimes/RB battery/NN can/VB get/VB very/RB hot/VB. |

Table 4. Pattern for extracting opinionated phrases by Htay & Lynn (2013)

| First Word | Second Word | Third Word |
|------------|-------------|------------|
| JJ         | NN /NNS     | -          |
| JJ         | NN /NNS     | NN /NNS    |
| RB/RBR/RBS | JJ/RB/RBR/RBS | NN/NNS    |
| VBN/VBD    | NN/NNS      | -          |
| RB/RBR/RBS | JJ          | -          |
| RB/RBR/RBS | VBN/VBD     | -          |
| RB/RBR/RBS | JJ          | -          |
| VBN/VBD    | RB/RBR/RBS  | -          |

Words[] = line.split(" ")
while words in Words[] do
    word = ReplaceSlangWord(word)
end while
postaggedline = POSTagger(line)
aspects = extract_aspect(postaggedline)
end for
output(aspect, Sentence[])
end function
function Reduce(aspect, Sentence[])
    ACount = count (aspects);
    SCount = count(sentences)
    Min Support = Scount/ACount
    for all sentence[] of aspect do
        if(count(sentence) > min Support)
            output (candidate aspect,Sentence[])
    end if
end for

After the aspect extraction, sentences are categorized as per the aspects. This result is given as input to next MapReduce Stage which does sentiment classification.

4.5. Sentiment classification
In the proposed work, Sentiment analysis is carried out using pattern-based Lexicon method. After the aspect extraction, we observed that sentiment pattern appears before and after the aspect patterns. These sentiment patterns are extracted, and Lexicon dictionary is used to classify the reviews. The Lexicon dictionary named SentiWordNet contains precalculated value for each word which can be used to predict the sentiment orientation of that word. The review sentences containing sentiment words are considered for sentiment classification. These sentences are classified into positive or negative sentences based on the sentiment score of sentiment words. The positive sentences usually contain the sentiment words such as good, awesome, super, and so on and negative sentences contain the sentiment words such as bad, worst, poor, and so on. The review sentences are initially tagged using Stanford POS tagger. Lexicon-based approach is applied for both uni-gram and bi-gram words.

4.5.1. Uni-gram method
The words tagged only as adjectives or adverbs are considered in ni-gram. These sentiment words are mapped against SentiWordNet to check the polarity. The sentiment score of the review sentence is calculated by finding the average sentiment score which is shown below.

\[
\text{Average Sentiment Score} = \frac{\text{Sum of Scores of Sentiment Words}}{\text{Total number of sentiment words}}
\]  

(1.1)

Based on the average sentiment score shown in Formula 1.1, the review sentences that contain sentiment word are classified as positive or negative.

4.5.2. Proposed bi-gram approach for sentiment analysis
Since unigram Lexicon method is built for uni-gram sentiment pattern, it gives misclassification for the bigram sentiment word. In the sentence, “This is awesome/JJ and Beautiful/NN,” only “awesome” is taken into consideration. Also, in uni-gram Lexicon method, sentiment words preceded by Negation words and adverbs are not taken for analysis. For example, “This phone is not good,” only “good” is taken for consideration and classifies sentence into positive; but “not” which is an adverb, gives negative effect on sentiment. Htay has proposed common pattern for aspect-based opinionated phrases. To get better results, in our proposed approach, we have written separate pattern for aspect extraction and opinion phrase extraction. The pattern used for opinion phrase extraction is shown in Table 5.

The proposed pattern extracts unigram opinion words which comes with noun phrases (first three patterns) and bi-gram opinion phrases (last four patterns).
Along with bi-gram pattern, Negation rule is applied to sentiment word preceded with negation word like “not.”

(a) If the sentiment word with positive score is preceded by “not,” then it is treated as Negative word (e.g. not good).
(b) If the sentiment word with Negative score is preceded by “not,” then it is treated as Positive word (e.g. not bad).

The proposed method for negation handling and sentiment classification is given in algorithm 1.2.

**Algorithm 1.2: Sentiment classification**

**Data:** dataset: = Real-time reviews, Sentiment word corpus  
**Result:** Classification: = Positive and Negative  
**Notations:** RW: review, SC: Sentiment Score, D: dataset, fw: first_word, sw: second_word  

```java
class Mapper  
function Map (aspect, review)  
SC = 0, count = 0, opp = 0, opn = 0;  
while RW ∈ D do  
for all aspect ∈ RW do  
    count = 0;  
    for all bigrams in RW do  
        if (fw is JJ && (sw is NN || sw is JJ)) then  
            B← add bigram to the list //Extract the bigram  
            Count++;  
        else if (fw is RB && (sw is NN || sw is JJ || sw is VB)) then  
            B← add bigram to the list //Extract the bigram  
            Count++;  
        else if (fw is VB && (sw is NN|| sw is RB)) then  
            B← add bigram to the list //Extract the bigram  
            Count++;  
        end for  
        if (count > 0) then  
            SC = 0, rwp = 0  
            while words in B do  
                if (fw is (RB && (sw is NN || sw is JJ) && sw is VB)) then  
                    rwp++;  
                end if
            end while
        end if
    end for
end for
```

### Table 5. Pattern for opinion phrase extraction

| First Word | Second Word |
|------------|-------------|
| JJ         | NN /NNS     |
| RB/RBR/RBS | NN/NNS      |
| VBN/VBD    | NN/NNS      |
| RB/RBR/RBS | JJ          |
| JJ         | JJ          |
| RB/RBR/RBS | VBN/VBD     |
| VBN/VBD    | RB/RBR/RBS  |

Rodrigues et al., Cogent Engineering (2020), 7: 1810862  
https://doi.org/10.1080/23311916.2020.1810862
Get the sentiment score from SentiWordNet

if \( fw = \text{“not”} \) then
    \[ SC = (\text{score of } sw) \times (-1) \]
else
    \[ SC = \text{score of } fw + \text{score of } sw \]
end

end while

SC = \text{sum} (SW \in B)
AR = \frac{SC}{\text{rwp}}

end if

if AR>0 then
    opp = positive
elseif AR<0 then
    opn = negative
end if

end for

output (aspect, [opp,opn])

end
class Reducer

function Reduce (aspect, [opp,opn])
    int sum1 = 0, sum2 = 0;
    sum1 = sum1 + opp;
    sum2 = sum2 + opn;
    Output (aspect, [sum1, sum2])
end

The results are discussed in section 5.

5. Results and discussions

The proposed sentiment analysis is performed on Hadoop MapReduce framework. Apache Hadoop is installed on Ubuntu 16.04 with 8GB RAM. The Multi-Node setup consists of 4 nodes with one node as master and remaining three nodes as slaves. The reviews of electronic gadgets are fetched from amazon.com with the help of a Java HTML parser Jsoup and are stored in HDFS. The electronic gadgets considered for experimental purpose include Nokia 6610, Apex DVD player, Canon g3, Nikon Coolpix, Creative labs jukebox, and Canon S100. MapReduce programming model is used for the further analysis of extracted reviews.

5.1. Datasets

The six sets of product reviews are generated out of the extracted reviews from amazon.com and review datasets from Hu and Liu research (Hu & Liu, 2004). Each review set has different number of reviews and aspects. The number of reviews considered for each product and their annotated aspects is shown in the Table 6.
5.2. Aspect extraction
We have proposed trigram pattern for explicit aspect extraction. Out of the extracted aspects, there are several nonaspect phrases listed in Table 7. To extract frequent features, we have used Apriori algorithm for filtering.

Table 8 shows the most frequent features spoken by customer after filtering process.

Accuracy is calculated for the aspects before and after filtering process according to the formula given below:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

(1.2)

| Dataset           | Reviews | Number of aspects |
|-------------------|---------|-------------------|
| Canon G3          | 642     | 105               |
| Jukebox           | 1810    | 178               |
| Nikon Coolpix     | 380     | 74                |
| Nokia-6610        | 588     | 96                |
| Canon S100        | 349     | 88                |
| Apex DVD Player   | 838     | 104               |

Table 7. Nonaspect phrases

| Patterns             | Phrases                                                   |
|----------------------|-----------------------------------------------------------|
| DT NN/NNS NN/NNS     | the/DT family/NN plan/NN                                  |
| JJ NN/NNS NN/NNS     | Recent/JJ road/NN trip, tid/JJ bit/NN quieter/NN          |
| DT NN/NNS VBZ/VBP/VB  | that/DT glitters/NNS is/VBZ, any/DT disadvantages/NN are/VB |
| RB NN/NNS VBZ/VBP/VB  | ok/RB everybody/NN loves/VBZ                              |
| NN NN/NNS NN/NNS     | tech/NN help/NN person/NN, currency/NN length/NN area/NN  |
| DT NN/NNS            | a/DTD blessing/NN, no/DTD way/NN                         |
| JJ NN/NNS            | decent/JJ things/NNS, big/JJ mechanism/NN                |

Table 8. Extracted aspects

| Product              | Features spoken by customers more often (Extracted aspects) after filtering |
|----------------------|----------------------------------------------------------------------------|
| Nokia                | Battery life, camera, color screen, feature, keys, volume, voice dialing, phone, screen, speaker phone, radio |
| Canon                | Camera, flash, battery, image quality, options, shutter, price, menu, resolution, zoom, control |
| Nikon                | Lens, card reader, memory card, camera, manual, image quality, mode, speed, settings |
| Jukebox              | Storage capacity, tech support, playlist, navigation, scroll wheel, music, volume, warranty |
| Apex DVD             | Button, customer service, dvd player, control, disk, feature, replacement, model, manual |
Table 9. Aspect extraction results

| Dataset       | Predicted aspects using proposed method (before filtering) | Predicted aspects (after filtering) | Accuracy in % (after filtering) |
|---------------|-------------------------------------------------------------|-------------------------------------|---------------------------------|
| Canon G3      | 183                                                         | 92                                  | 60                              |
| Jukebox       | 305                                                         | 124                                 | 59                              |
| Nikon Coolpix | 156                                                         | 59                                  | 63                              |
| Nokia-6610    | 215                                                         | 78                                  | 58                              |
| Canon S100    | 188                                                         | 91                                  | 57                              |
| Apex DVD Player| 203                                                         | 83                                  | 61                              |

T_P is a case in which a word is an aspect and it is predicted as aspect, T_N is a case in which a word is not an aspect and it is not predicted as aspect, F_P is a case in which a word is not an aspect but it is predicted as aspect, and F_N is a case in which a word is an aspect but not predicted as aspect. The Table 9 shows the aspect extraction results.

For our annotated dataset, the experimental result shows that the accuracy of aspect extraction using proposed trigram pattern is in the range of 29 ~ 42%, but trigram pattern with filtering process outperforms with the accuracy 57 ~ 63%.

5.2.1. Sentiment classification
Sentiment classification of particular aspect is performed using Lexicon-based approach with uni-gram and bi-gram pattern-based method. Since the analysis is performed at aspect level, some of the aspects extracted for mobile products are Volume, Battery Life, and Camera. The Figure 6 depicts the result of classification of reviews based on aspect for a product Nokia. The result is shown among nine most frequently spoken aspects of Nokia reviews.

5.2.2. Performance evaluation of sentiment classification methods
The performance of sentiment classification is evaluated using performance measures, namely, Precision, Recall and F1-measure. The experiment showed that, proposed bi-gram method shows the significant increase in the precision and recall compared to uni-gram method and Htai pattern. In dataset Canon S100, Nikon, Apex and Jukebox, Apex, Canon G3, and Nokia the Precision rate increased in the range of 21~33% compared to unigram and 3~5% when compared...
Similarly, there is an increase in recall rate compared to other two approaches in the range of 5%. Table 10 and 11 shows the Precision and Recall of the experiments.

The experimental result showed that the Lexicon uni-gram method gives accuracy ranging 53 ~ 55%, Htay pattern results in accuracy in the range 66~70% but bi-gram lexicon method outperforms with the accuracy 72 ~ 75%. The value of the F1-measure has increased on average 4% compared to Htay pattern. Figure 7 shows the performance measures of sentiment classification methods.

Figure 8 depicts difference between processing time in normal setup and in Hadoop, from which it can be concluded that the aspect extraction process is faster in Hadoop cluster compared to normal setup.

6. Conclusion
The popularity of online shopping has increased in much consideration in the zone of sentiment analysis. The proposed work concentrates on performing aspect level sentiment analysis of product reviews. This work focused on extracting reviews from amazon.com, preprocessing, POS tagging, comparing different aspect selection methods and sentiment classification. The experimental result proves that proposed pattern-based aspect extraction gives better results with filtering process. Sentiment classification result concludes that, proposed bi-gram method has gained high accuracy when compared to uni-gram method and Htay pattern. The entire experiment is conducted on Hadoop Single Node and Multi-Node cluster to compare the execution time of both the clusters. From the outcome it’s been proved that Hadoop Multi-Node cluster performance is faster when compared to Single Node.
The presented work focuses only on explicit aspects. Identification of implicit aspects can be considered in further process. In the future work, it is also possible to improve the proposed method by considering pronoun resolution, sarcastic review sentences with different sentiment classification techniques.

Funding
The authors received no direct funding for this research.

Author details
Anisha P. Rodrigues
E-mail: anishapr@nitte.edu.in
Niranjan N. Chiplunkar
Roshan Fernandes
E-mail: roshan_nmamit@nitte.edu.in
1 Department of Computer Science and Engineering, NMAM Institute of Technology, Nitte, India.

Citation information
Cite this article as: Aspect-based classification of product reviews using Hadoop framework, Anisha P. Rodrigues, Niranjan N. Chiplunkar & Roshan Fernandes, Cogent Engineering (2020), 7: 1810862.

References
Bafna, K., & Toshniwal, D. (2013). Feature based summarization of customers' reviews of online products.

Procedia Computer Science, 22, 142–151. https://doi.org/10.1016/j.procs.2013.09.090
Chauhan, N., & Singh, P. (2017, March). Feature based opinion summarization of online product reviews. In 2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM) (pp. 1–7). IEEE.
Chifu, E. Ş., Leţia, T. Ş., & Chifu, V. R. (2015, October). Unsupervised aspect level sentiment analysis using Ant Clustering and Self-organizing Maps. In 2015 International Conference on Speech Technology and Human-Computer Dialogue (SpDeD) (pp. 1–9). IEEE.
Deshmukh, J. S., & Tripathy, A. K. (2018). Entropy based classifier for cross-domain opinion mining. Applied Computing and Informatics, 14(1), 55–64. https://doi.org/10.1016/j.aci.2017.03.001
Devasia, N., & Shelk, R. (2016, October). Feature extracted sentiment analysis of customer product reviews. In 2016 International Conference on Emerging Technological Trends (ICETT) (pp. 1–6). IEEE.
Esparza, G. G., de-Luno, A., Zezzatti, A. O., Hernandez, A., Ponce, J., Alvarez, M., ... de Jesus Nova, J. (2017, June). A sentiment analysis model to analyze students reviews of teacher performance using support vector machines. In International Symposium on Distributed Computing and Artificial Intelligence (pp. 157–164). Springer.

Fang, X., & Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, 2(1), 5. https://doi.org/10.1186/s40537-015-0015-2

Gadekalli, T., Soni, A., Sarkan, D., & Kuruva, L. (2019). Application of sentiment analysis in movie reviews. In Sentiment Analysis and Knowledge Discovery in Contemporary Business (pp. 77–90). IGI Global.

Ghosh, M., & Sanyal, G. (2017). Preprocessing and feature selection approach for efficient sentiment analysis on product reviews. In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (pp. 721–730). Springer.

Guo, K., Shi, L., Ye, W., & Li, X. (2014, May). A survey of internet public opinion mining. In 2014 IEEE International Conference on Progress in Informatics and Computing. https://nip.stanford.edu/software/tagger.shtml

Htay, S. S., & Lynn, K. T. (2013). Extracting product features and opinion words using pattern knowledge in customer reviews. The Scientific World Journal, 2013, 1–5. Hindawi Publishing Corporation. https://doi.org/10.1155/2013/394758.

Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168–177).

Jinturkar, M., & Gotmare, P. Sentiment analysis of customer review data using big data.

Kang, D., & Park, Y. (2014). based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach. Expert Systems with Applications, 41(4), 1041–1050. https://doi.org/10.1016/j.eswa.2013.07.101

Khan, F. H., Qamar, U., & Boshir, S. (2017). A semi-supervised approach to sentiment analysis using revised sentiment strength based on SentiWordNet. Knowledge and Information Systems, 51(3), 851–872. https://doi.org/10.1007/s10115-016-0993-1

Kumar, K. S., Desai, J., & Majumdar, J. (2016, December). Opinion mining and sentiment analysis on online customer review. In 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC) (pp. 1–4). IEEE.

Li, Y., Qin, Z., Xu, W., & Guo, J. (2015). A holistic model of mining product aspects and associated sentiments from online reviews. Multimedia Tools and Applications, 74(23), 10177–10194. https://doi.org/10.1007/s11042-014-2158-0

Maddikunta, P. K. R., Gadekalli, T. R., Al-Ahmari, A., & Abidi, M. H. (2020). Location based business recommendation using spatial demand. Sustainability, 12(10), 4124. https://doi.org/10.3390/su12104124

Maharani, W., Widiantora, D. H., & Khodra, M. L. (2015). Aspect extraction in customer reviews using syntactic pattern. Procedia Computer Science, 59, 244–253. https://doi.org/10.1016/j.procs.2015.07.545

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Moraes, R., Vellati, J. F., & Neto, W. P. G. (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. Expert Systems with Applications, 40(2), 621–633. https://doi.org/10.1016/j.eswa.2012.07.059

Parlar, T., & Özel, S. A. (2016, August). A new feature selection method for sentiment analysis of Turkish reviews. In 2016 International Symposium on Innovations in Intelligent Systems and Applications (INISTA) (pp. 1–6). IEEE.

Pottar, A., Patil, P., Bagla, R., Pandey, R., & Jadhav, N. (2016). SAMIKSHA-sentiment based product review analysis system. Procedia Computer Science, 78, 513–520. https://doi.org/10.1016/j.procs.2016.02.096

Reddy, G. T., Reddy, M. P. K., Lakshmamma, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. IEEE Access, 8, 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942

Sharma, A., & Dey, S. (2012, October). A comparative study of feature selection and machine learning techniques for sentiment analysis. In Proceedings of the 2012 ACM research in applied computation symposium (pp. 1–7).

Subramaniyaswamy, V., Vijayakumar, V., Logesh, R., & Indragandhi, V. (2015). Unstructured data analysis on big data using map reduce. Procedia Computer Science, 50, 456–465. https://doi.org/10.1016/j.procs.2015.04.015

White, T. (2012). Hadoop: The definitive guide. O’Reilly Media, Inc.

Yang, Y., Xu, C., & Ren, G. (2012). Sentiment analysis of text using SVM. In Electrical, Information Engineering and Mechatronics 2011 (pp. 1133–1139). Springer.

Zhang, L., Hua, K., Wang, H., Qian, G., & Zhang, L. (2014). Sentiment analysis on reviews of mobile users. Procedia Computer Science, 34, 458–465. https://doi.org/10.1016/j.procs.2014.07.013
