Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 2): pp. 827-836
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.02.93

Sentiment Analysis of Online Food Reviews using Big Data Analytics

Hafiz Muhammad Ahmed, Department of Software Engineering, University of Management and Technology, Lahore, Pakistan.
Mazhar Javed Awan, Department of Software Engineering, University of Management and Technology, Lahore, Pakistan, mazhar.awan@umt.edu.pk
Nabeel Sabir Khan, Department of Computer Science, University of Management and Technology, Lahore, Pakistan.
Awais Yasin, Department of Computer Engineering, National University of Technology, Islamabad, Pakistan
Hafiz Muhammad Faisal Shehzad, Department of Computer Science and IT, University of Sargodha, Sargodha, Pakistan

Abstract— Nowadays sentiment analysis has become very important, mostly used for huge datasets and helpful for researchers for applying methods and techniques. Amazon's food data is growing exponentially and traditional systems are unable to process it, so we used Big Data to overcome this problem. In this paper, we explore different methods and techniques of sentiment analysis using apache spark data processing system for big datasets of Amazon Fine Food reviews. Three mechanisms are applied that have more than 80% accuracy named as Linear SVC, Logistic Regression, and Naïve Bayes by using MLlib which is Apache Spark's library for ML. When applied these methods we realize that Linear SVC performs efficiently than NB and logistic regression.

Keywords— Sentiment Analysis; Apache Spark; reviews, Machine Learning, Big Data, Analytics

I. INTRODUCTION

Online reviews for your business are one of the most important elements in the case of marketing analytics. In business, online customer reviews become very important for products and services, therefore, we can trace out the bad and good reviews by the help of which we can analyze the product quality and their standards, also useful in making new methods and techniques for improving the quality of products [1]. Customer reviews are about the feedback as it contains a huge amount of data that is widely spreading every second and it contains the structured, unstructured, and sentiments' data called big data analysis and information extraction. The success of a company or product directly depends on its customer's feedback [2].

Sentiment analysis in the field of information retrieval computationally identifying and categorizing opinions from a piece of text in the form of positive or negative. For the huge amount of data, it becomes very crucial to analyze and use sufficient data for their functionality and therefore it's now a very big task for data warehouse and the relational database [3]. Big data analysis has three important aspects named velocity, volume, and variety. Collecting the huge size of data from various data contents at a specific time is referred to as volume. Velocity is the strength of data by which the data can be measured [4]. Various kinds of data having different sources including structured and unstructured data like text, audios, videos, and images' data recognized as a variety of data [5]. Most of the robotics, the complex machine learning mechanisms, and techniques used for big data's requirements. As the data amount is huge, not capable of being used in a personal computer's memory, therefore some tools of machine learning like R and Weka can be used for big data analysis, Some new tools are now introduced such as apache spark apache Hadoop, these tools can easily handle the machine learning algorithms and work very efficiently and can acquire high rate performance [6].

Apache Spark was introduced by the University of California in 2009. It can be used mostly for high-size data, but it can work efficiently for both the batch and streaming data, also easily handle APIs on huge datasets [7]. Spark is a most efficient framework for large data than other optimizations like Hadoop and achieves high performance. Spark MLlib is a scalable library and can also be used with different high-level programming languages [8]. The application of a spark comprises 5 basic entities that are shown in Fig. 01.

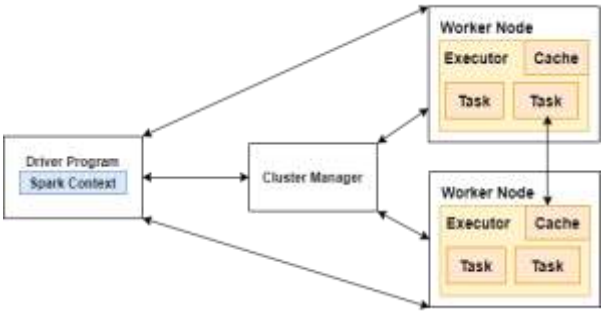

Fig. 01. Basic 5 entities of a spark application

The generation of fake news, comments, and sentiments is growing daily from websites and social media [9]. Sentimental analysis is used to resolve various machine learning problems [10]. In this paper, we used the spark MLlib as it is a modern library that was created in 2014. Research has been made by using spark's MLlib and it can be helpful for big data analysis.
Very limited research was carried out on spark MLlib for large size data but much work requires in this field. By using various techniques and methods on spark's MLlib, explore new different ways for analysis of big data to achieve high efficiency and performance.
As per our knowledge, our work is the first on sentiment reviews related to amazon food on big data using machine learning

II. RELATED WORK

Scientists are actively researching sentiment analysis that has become the biggest area of research in the last few years. Sultana, Kumar [11] described that sentimental analysis has three important aspects, positive, negative, and neutral. From last few years, the world wide web becomes a key factor of customers' reviews, by the social media and e-commerce websites, such as Facebook, tweeters user can share their reviews and these reviews can be good or bad, and these reviews help in making choices about applying new plan and decisions about products. Chen, Xue [12] introduced a new technique to remove the traits of sentiment analysis for the reviews of products. The most common TF-IDF vectors can obtain by using the same form of synonyms by viewing the products' reviews, we can categorize the sequences of feature vectors along with clustering algorithms. By applying this technique we can refine span algorithms for pseudo consecutive phrases with FPCD having word order details. By using the last steps, the text feature is gathered. As a result of applying the different mechanisms of performance can be enhanced. In Abbas, Memon [13], the authors introduced a new heuristic method along with naïve bias for specified issues. An MNB is an NB classifier used for text categorization and implemented for sentimental analysis. The results of high data references verify the efficiency of used algorithms. In Neethu and Rajasree [14], the authors examined Twitter posts by using ML techniques for different products like a mobile, pad, and laptop, etc. these strategies applied for Twitter sentiment analysis. Using sentimental analysis it is easy to explore the main consequences in sentiment analysis. Some issues can create and for resolving these issues feature extraction can do after preprocessing in two steps. In the first step, features are firstly removed from tweets and then done features extraction and then added to feature vector. Feature classification is done by applying classifiers like NB, SVM, and maximum entropy. Indra, Wikarsa [15] introduced web-based applications that categorize tweets of netizens into four different types of machine learning algorithms applied named logistics regression. Mainly four different types of methods are used for extracting tweets, text features, and machine learning methods. In this work total of 1800 tweets were used as a training dataset. Various techniques can apply in real-time processing such as the transfer of URL, punctuation, stop words, tokenization, and stemming. The group of features used for logistic regression techniques for classification. Obtain high-efficiency tweets is about 92% by applying a confusion matrix. In Wawre, Deshmukh [16], the authors applied different sentiment analysis methods to movie review. They compared two different techniques like SVM and Naïve Bayes by performing these two techniques to movie reviews and conclude that Naïve Bayes is most efficient and performs well than SVM. If we apply Naïve Bayes on training datasets of a huge number of reviews then it can give the best and accurate results. In Laksono, Sungkono [17], the authors categorized customer reviews from the Trip Advisor in the best restaurant of port city Surabaya. Also, we determine the method of exploring restaurant customer reviews by performing and differentiate both Naïve Bayes and Text Blob. These two techniques perform well, but Naïve Bayes gives the most efficient and valid result than Text Blob. Liu,

Blasch [18] used and scale-out naïve Bayes classifiers for huge datasets. Applying NBC for exploration mechanisms for the best performance. The big data analysis system used for this work. By this process, NBC's precision level increased and then gain 82%. We can say NBC is used to explore views of reviews with high productivity. Prabhat and Khullar [19] performed assessment of the SA on the reviews of customers by adopting the Apache Spark. Apache Spark, which is an extensible framework is used, and different methods of MLlib are applied as like Naïve Bayes, SVM and logistic regression. By applying these methods we concluded that SVM works well and more precise than Naïve Bayes and logistic regression and gives the most accurate results. The recent studies in year 2021 Mazhar and Shafry [20, 21] used big data with Spark ML framework to predict stock market and black Friday sales in an efficient way.

III. DATA SET AND METHODOLOGY

The approach which is used in this paper has five stages that are shown in Fig. 02 These stages are acquiring dataset through various visualization, data preprocessing, extraction of features implementing the machine learning classifiers through Spark MLlib, and lastly, evaluated models through train test split by using different metrics of binary classification.

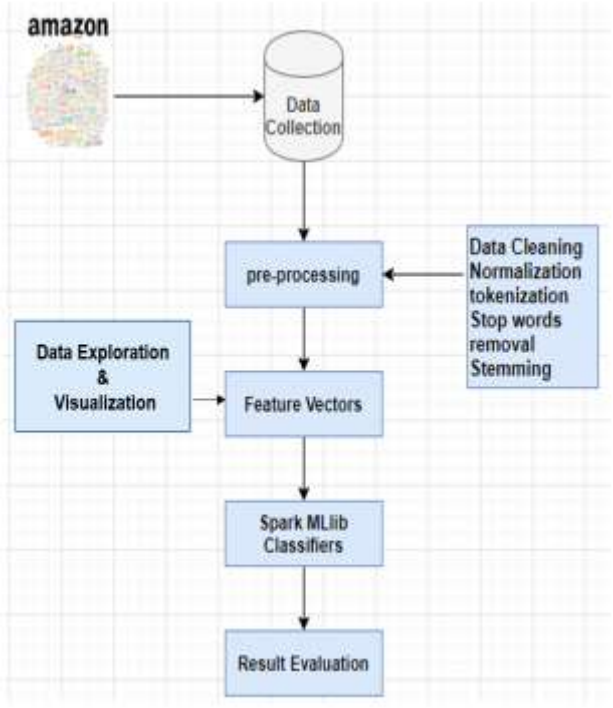

Fig. 02. Stages of the proposed approach

A. Dataset

Amazon's Fine Food dataset [22] is used for experiments. Amazon's dataset consists of 568454 reviews, the number of users is 256,059, several products are 74,258, and the number of columns is 10. The features are unique product_id, unique user-id, profile name, number of users who found the review helpfulness numerator, number of users who indicated whether the review helpful or not, the score is based upon the rating 1 to 5, timestamp of the review, a summary of the review and text of the review.

B. Preprocessing Stage

Before training the models the dataset is passed into the preprocessing stage to provide the best input to the models for training. There are the following steps that are necessary cleaning of data also called the data wrangling stage.

- Firstly, all the null and duplicated values are identified and removed from the text.
- Secondly, remove the noise from text data by removing irrelevant data which can lead to reducing the performance of classifiers like non-alphabetic characters, digits, special characters and punctuation marks.

➢ Thirdly, normalize the score column that has 1 to 5 values shown in Fig. 03.

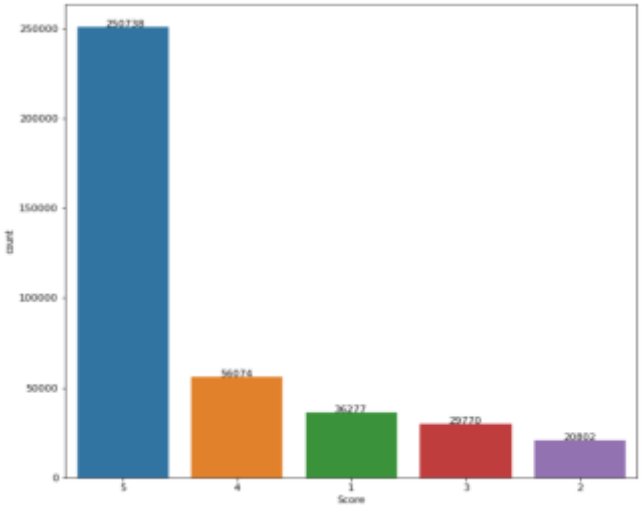

Fig. 03. Distribution of score of review from 1 to 5

For this, generated a new column named as a label which has 0 or 1 value based upon the helpfulness features as shown in fig 04.

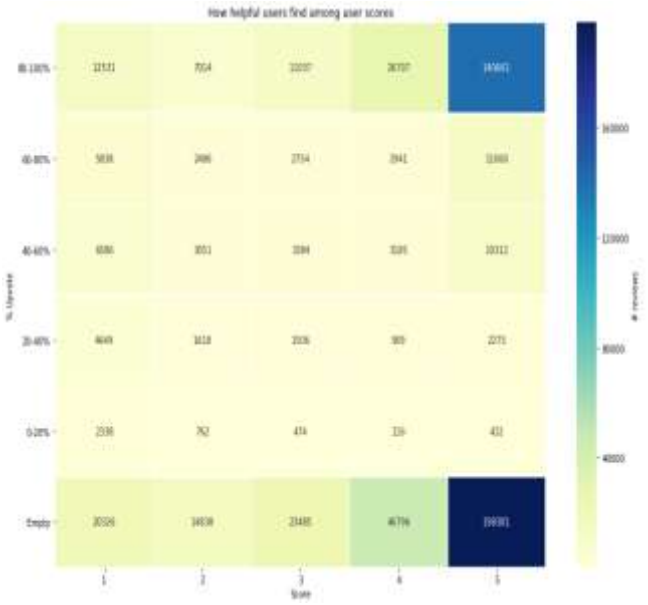

Fig. 04. Helpfulness review against score

➢ Lastly, tokenize the preprocessed text on the basis of space, and remove the stop words that are the common words and have little significant in the sentence like preposition and conjunction in the table 01.

Table.01: Preprocessing tokenization and stop words removal

Preprocessing	**sample comment**
Without tokenizing	'yesterday i ordered food from kfc the food was not not cooked properly the taste of the food was very bad it was the 2nd time i faced the same problem'
After tokenizing	'yesterday', 'i', 'ordered', 'food', 'from', 'kfc', 'the', 'food', 'was', 'not', 'not', 'cooked', 'properly', 'the', 'taste', 'of', 'the', 'food', 'was', 'very', 'bad', 'it', 'was', 'the', '2nd', 'time', 'i', 'faced', 'the', 'same', 'problem'

Before stop words	'yesterday', 'i', 'ordered', 'food', 'from', 'kfc', 'the', 'food', 'was', 'not', 'not', 'cooked', 'properly', 'the', 'taste', 'of', 'the', 'food', 'was', 'very', 'bad', 'it', 'was', 'the', '2nd', 'time', 'i', 'faced', 'the', 'same', 'problem'
After stop words	'taste', 'ordered', 'faced', 'time', 'problem', 'bad', 'properly', 'cooked', 'yesterday', 'food', 'kfc', '2nd'

C. Data Exploration

Data exploration and visualization is an important stage for analytics. For this, we used matplotlib and seaborn libraries to explore data.
Fig. 05 shows the correlation matrix between the five features.

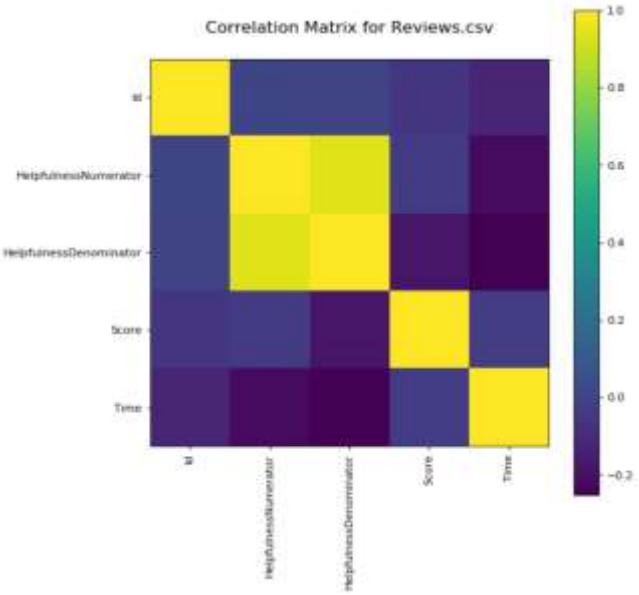

Fig. 05. Correlation Matrix of five features

The summary feature is very important in which review is present, in fig 06 shows the most frequent terms which are love, good, best, taste, delicious and yummy.

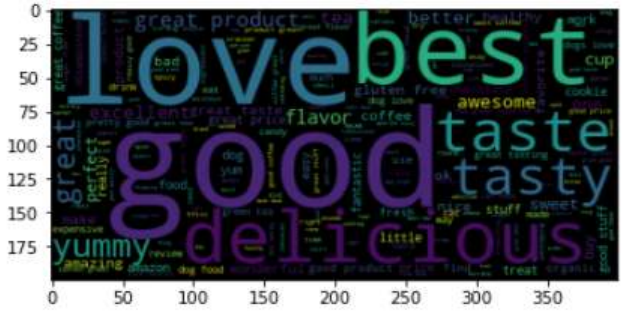

Fig. 06. Most frequent terms in the reviews

Fig. 07 shows the frequent terms against the low scoring that is 1 in the review. The terms are garbage, price, horrible, bad nasty, rip tasteless and warning are negative sentiments with a score of 1.

Fig. 07. Frequent terms against low scoring in the reviews

Fig. 08 shows the frequent terms against the high scoring that is 5 in the review. The terms are great, good, oatmeal, spice, tasty, delicious, and food are positive sentiments with a score of 5.

Fig. 08. Frequent terms against high scoring in the reviews

D. Feature extraction

Feature engineering is another important stage before training of the model. Based on the section of data exploration in which fig. 04 shows that the reviews are skewed towards positive, half of the reviews are with zero vote and many people agree with the highest score. For this, we converted the score into two classes, positive and negative label as shown in fig.09.

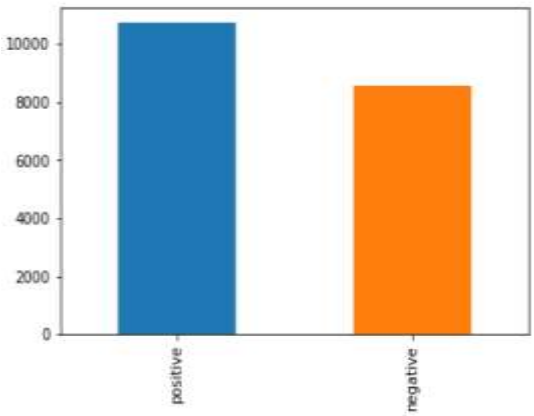

Fig.09. count of the class label after normalized

Text data is transformed into feature vectors before training the models so that models can be trained efficiently. TF-IDF is used in feature extraction. TF counts the words that occur in the text and IDF counts how common and rare words present in all documents or text.

E. SparksML Classfiers models

Different Classifiers are trained and evaluated but 3 classifiers that have more than 80% accuracy are selected in this paper. Logistic Regression (LR), Linear Support Vector Classifier (LinearSVC), and Naïve Bayes (NB).

LinearSVC: The most appropriate machine learning method is LinearSVC. This is mainly a technique for classifying linear issues. The aim of a LinearSVC (Support Vector Classifier) is to fix the data give back the best hyperplane that classifies the data. After the hyperplane, for prophecy, we have to use some characteristics. This makes it possible for using the specific mechanisms and techniques in any way. LinearSVC has adaptable execution of SVC with the linear kernel. In the sklearn testimonial, the process used in LinearSVC is more accurate [23].

Naïve Bayes (NB): is Bayesian theorem based on the supervised machine learning technique which is used for classification tasks. The Bayesian theorem is explained following:

$$P(a/b) = \frac{P(a)P(b/a)}{P(b)} \quad (1)$$

"a" and "b" are independent events. P (*a/b*) is a conditional probability which means that the probability of '*a'* happened when event 'b' has occurred. *P (b/a)* is the probability of event '*b*' when '*a*' has occurred. *P (a), P (b*) both are the probabilities of the events '*a'* and '*b'*. Naïve Bayes build up the model by fixing the distribution of every feature [24, 25].

Logistic regression: is a machine learning algorithm that is broadly used by researchers. The logistic regression comprehends variables' vector and find out the coefficient for input expressions and then trace out the class of text as a word vector. The logistic regression function determines multiple linear functions expressed as

$$Logit\ (P) = \beta 0 + \beta 1X1 + \beta 2X2 \ldots . \beta kXk \quad (2)$$

P represents the probability of the occurrence of the feature. X1, X2…. Xk represents the value of predictor and B1β2 …. Bk represents the model's intercept [19, 26].

IV. EXPERIMENT SETUP

We used the data bricks Spark cloud with python 3.0. The 82007 negative and 486447 positive samples are present in our classes. To balance the classes from the total dataset 82007 samples for the positive class and the same size of samples for the negative class are taken for experiments. By randomly splitting 80 % for training and 20 % for testing from the total.Fig. 10 shows the summary of the dataset by positive and negative class labels.

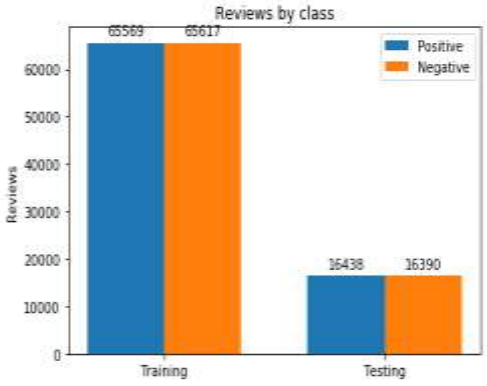

Fig. 10 Bar chart shows the reviews training and testing data.

V. EXPERIMENTAL RESULTS

The basic aim of experimental results is to check the classifiers' performance that performed using Spark MLlib such as LR, NB, and Linear SVC. The performance of all the classifiers is measured by using various evaluation metrics confusion matrix which is shown in table 02.

Table 2. The Results of classifiers using different evaluation metrics

Models	Accuracy	Precision	Recall	F1-Score
Logistic Regression	87.38	86.54	88.78	87.64
Naïve Bayes	83.43	82.35	88.78	85.44
LinearSVC	88.38	88.54	88.39	88.46

The fig.11 shows the accuracies of each model.

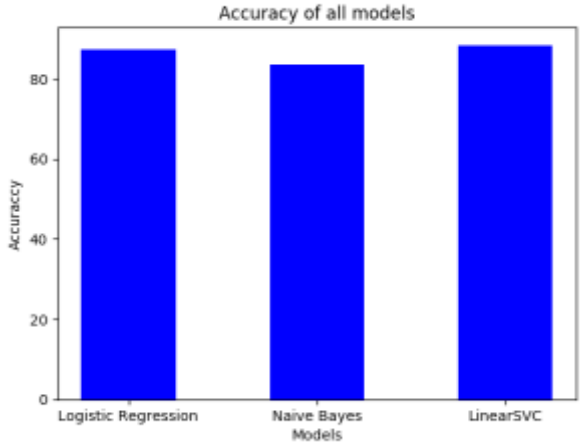

Fig. 11. Accuracy comparison of all classifiers

Fig. 12 shows the evaluation in the form of a bar chart.

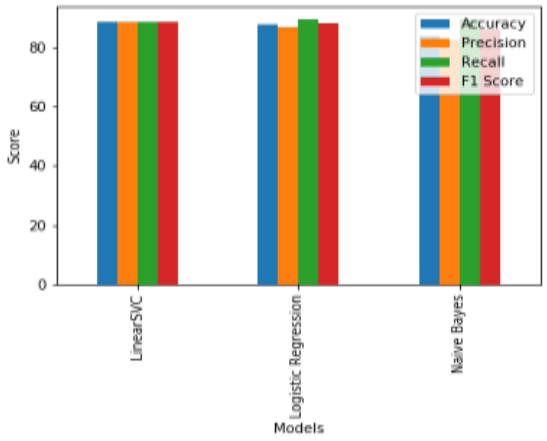

Fig. 12. Evaluation of classifiers using different metrics

Moreover, we also calculated the computation time in table 03 of training data and table 04 the computation time of test data.

Table 3. Running time comparison of training dataset in seconds

Execution Time of training dataset			
Dataset	Logistic Regression	Naïve Bayes	LinearSVC
130995	40.2	20.65	346.2

Table 4 Running time comparison of test dataset in seconds

Execution Time of testing dataset			
Dataset	Logistic Regression	Naïve Bayes	LinearSVC
33019	0.06	0.06	0.15

VI. CONCLUSION

Sentiment analysis is a process that identifies the people's emotions from text data and is commonly used in business for the betterment of products or services' quality. In this article, the sentiment analysis (SA) is implemented on Amazon's Fine Food Reviews dataset in a big data context. Different demonstrations were carried out for analyzing the sentiments having large datasets by applying various classification techniques like NB, Linear SVC, and LR by using Spark MLlib. For large datasets Apache spark MLlib is used. For the experiment 131186 reviews are used in model training and 32829 reviews are used for testing the models. Few steps were applied for exploring and analyzing the data. The algorithms like Naïve Bayes, LinearSVC, and logistic regression were applied. By performing the analysis it shows that the linear support vector classifier works well than other classifiers. In the future, to improve the performance of the classifier different features set will be considered such as bi-gram, tri-gram, and four-gram.

REFERENCES

1. Hamzah, A.A., M.F.J.J.o.U.S.S. Shamsudin, and Technology, *Why customer satisfaction is important to business?* Journal of Undergraduate Social Science and Technology, 2020. **1**(1).
2. Alam, T.M. and M.J. Awan, *Domain analysis of information extraction techniques.* International Journal of Multidisciplinary Sciences and Engineering, 2018. **9**: p. 1-9.
3. Rehman, A.A., M.J. Awan, and I. Butt, *Comparison and Evaluation of Information Retrieval Models.* VFAST Transactions on Software Engineering, 2018. **6**(1): p. 7-14.
4. Hajjaji, Y., et al., *Big data and IoT-based applications in smart environments: A systematic review.* 2021. **39**: p. 100318.
5. Sagiroglu, S. and D. Sinanc. *Big data: A review*. in *2013 international conference on collaboration technologies and systems (CTS)*. 2013. IEEE.
6. Zheng, J. and A. Dagnino. *An initial study of predictive machine learning analytics on large volumes of historical data for power system applications*. in *2014 IEEE International Conference on Big Data (Big Data)*. 2014. IEEE.
7. Zaharia, M., et al., *Apache spark: a unified engine for big data processing.* 2016. **59**(11): p. 56-65.
8. Meng, X., et al., *Mllib: Machine learning in apache spark.* 2016. **17**(1): p. 1235-1241.
9. Abdullah, et al., *Fake News Classification Bimodal using Convolutional Neural Network and Long Short-Term Memory.* International Journal on Emerging Technologies, 2020. **11**(5): p. 209-212.
10. Medhat, W., A. Hassan, and H.J.A.S.e.j. Korashy, *Sentiment analysis algorithms and applications: A survey.* 2014. **5**(4): p. 1093-1113.
11. Sultana, N., et al., *Sentiment Analysis for product review.* 2019. **9**(3).
12. Chen, X., et al., *A novel feature extraction methodology for sentiment analysis of product reviews.* 2019. **31**(10): p. 6625-6642.
13. Abbas, M., et al., *Multinomial Naive Bayes classification model for sentiment analysis.* 2019. **19**(3): p. 62.
14. Neethu, M. and R. Rajasree. *Sentiment analysis in twitter using machine learning techniques*. in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. 2013. IEEE.
15. Indra, S., L. Wikarsa, and R. Turang. *Using logistic regression method to classify tweets into the selected topics*. in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. 2016. IEEE.
16. Wawre, S.V., S.N.J.I.J.o.S. Deshmukh, and Research, *Sentiment classification using machine learning techniques.* 2016. **5**(4): p. 819-821.

17. Laksono, R.A., et al. *Sentiment Analysis of Restaurant Customer Reviews on TripAdvisor using Naïve Bayes*. in *2019 12th International Conference on Information & Communication Technology and System (ICTS)*. 2019. IEEE.
18. Liu, B., et al. *Scalable sentiment classification for big data analysis using naive bayes classifier*. in *2013 IEEE international conference on big data*. 2013. IEEE.
19. Prabhat, A. and V. Khullar. *Sentiment classification on big data using Naïve Bayes and logistic regression*. in *2017 International Conference on Computer Communication and Informatics (ICCCI)*. 2017. IEEE.
20. Awan, M.-J., et al., *Social Media and Stock Market Prediction: A Big Data Approach.* Computers, Materials and Continua, 2021. **67**(2): p. 2569--2583.
21. Awan, M.-J., et al., *A Big Data Approach to Black Friday Sales.* Intelligent Automation & Soft Computing, 2021. **27**(3): p. 785--797.
22. McAuley, J.J. and J. Leskovec. *From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews*. in *Proceedings of the 22nd international conference on World Wide Web*. 2013.
23. Awan, M.J., et al., *Efficient Detection of Knee Anterior Cruciate Ligament from Magnetic Resonance Imaging Using Deep Learning Approach.* Diagnostics (Basel), 2021. **11**(1).
24. Al-Saqqa, S., G. Al-Naymat, and A.J.P.C.S. Awajan, *A large-scale sentiment data classification for online reviews under apache spark.* 2018. **141**: p. 183-189.
25. Awan, M.J., *Acceleration of Knee MRI Cancellous bone Classification on Google Colaboratory using Convolutional Neural Network.* International Journal of Advanced Trends in Computer Science and Engineering, 2019. **8**(1.6): p. 83-88.
26. Ali, Y., et al., *Detection of schistosomiasis factors using association rule mining.* IEEE Access, 2019. **7**: p. 186108-186114.