DATA MINING AND FEATURE EXTRACTION TECHNIQUES FOR OPINION MINING ANALYSIS IN TWITTER

A. AL FIRTHOUS[1], P.ARUL[2]

[1]Research Scholar
Bharathiar University
Coimbatore, Tamil Nadu, India
[2]Assistant Professor
Department of Computer Science and Applications
Government Arts College, Tiruchirapalli – 22
Tamil Nadu, India

E-mail: [1] alfirquest@gmail.com

ABSTRACT

By keeping and gaining customers, every small and large firm hopes to increase earnings and sales. It is vital to estimate future sales in order to improve sales. Aim of this work is to find the product is liked and recommended by the customer or not. To retain customers, product sellers must assess online customer demand in order to manage product, brand, and inventory levels. To improve customer service, online retailers or sellers must first understand the needs of their customers. It's also important to think about future sales. The proposed opinion mining on Amazon products is based on various sorts of real-time data obtained from Twitter. Lexical features, sentiment features and pragmatic features have also been used to improve the classification of review opinion features. After preprocessed the collected real time reviews have been involved in the process of feature extraction. Finally positive and negative opinions about the Amazon products have been classified by the proposed classification model. The proposed work findings and results states whether the Amazon shopping customers likes and recommend the product or not recommend. If the positive opinionated reviews are higher, then the customers like the product and recommend the product to purchase. If the negative opinionated reviews are higher, then the customers not like the product and not recommend the product to purchase. This research aims to aid clients in making purchasing decisions.

Keywords: *Opinion Mining, Rule Based, Lexicon, Pragmatic, Prediction, and Classification)*

1. INTRODUCTION

Online buyers must understand the best E-Commerce website to make purchases. But the problem is purchasing people cannot get the clear point whether the product is recommended or not by the previous buyers, because the commented and shared review's about the particular product are huge volume. The problem presented above could be approached by the proposed opinion mining process. Python and Natural Language Processing were used to collect and opinionated reviews about the product items, preprocess the collected dataset, further analysis, classification and visualize the data. This research looks into online reviews on social media product related information for the purpose of this opinion mining process. This work also provides a comparison of recent research on opinion mining approaches. That is, it evaluates and categorizes the positive and negative attitudes of purchasing customers toward the Amazon products categories. This research also aims to aid in making operational management decisions, such as determining product sales or customer product and service demand. The dataset for this study consists of online reviews of the product on social media Tweets.

Opinion mining can be defined as the process of extracting meaningful information from unstructured textual content from a variety of sources, such as posted or shared chat on Twitter, Facebook, and online blogs. Most of opinion mining process carried out by the traditional machine learning procedures. But the limitation of the previous suggestion of opinion mining task is not included much features especially the skip-gram features and retweets features. This limitation in opinion mining requires the creation of automated rule based systems that capable to analyze opinion data using a set of predefined rules. Opinion mining is also carried out by automated systems that employ any of the machine learning systems. In other cases, rule based and automated machine learning systems are integrated to create hybrid opinion mining model.

At the same time, opinion mining investigation is significant and complicated for analyzing a customer's purchasing experience and comments published across social media platforms. Natural language processing (NLP) is the most important approach for enhancing opinion mining. NLP is primarily interpreting the unstructured content into structuring in order to facilitate opinion mining. Opinion mining is the principal application area in which NLP has key responsibility.

2. RELATED WORKS

Hassani et al [16], have been recommended text mining is a tough approach for leveraging the potential of formless textual data to extract new information by analyzing it and expose noteworthy trends in big data analytics and hidden relationships in the data. This work assess the current status of text mining, opinion mining [6], [21], [22] study by reviewing modifies in published literature over the last few years and providing useful information practitioner and academic perspectives on the most common trends, methodologies, and applications of study of text mining. Additionally, the advantages and disadvantages of text mining are briefly discussed. Using diverse permutations of sanitizing procedures can help in enhancing text classification [14] results. As a result, various machine learning mechanisms investigated the impact of all possible grouping of five or six sanitizing procedures [3], [15] on four standard text corpora by extending text classification mechanisms. The suggestion of this work is that using a wide diverse sanitizing procedure [13] in grouping with text segregation is recommended to increase text classification accuracy [13].

Dalipi et al [10], have been suggested, Massive Open Online Courses (MOOCs), [5] sentiment analysis can be used to assess course notes, allowing mentor to effortlessly estimate their courses. This article is describes the use of sentiment analysis for investigate pupil's opinion in MOOCs, [8] with a focus on research papers published between January 1, 2015, and March 4, 2021. [2] Have been created a model that forecasts an public knowledge of preventive actions in Saudi Arabia's five primary regions. Dataset of Arabic COVID-19-related tweets have been collected during the pandemic [20] period for this investigation. The collected dataset has been evaluated by machine learning forecasting mechanisms, and the N-gram element mining mechanism namely [11] Nave Bayes (NB), K-nearest neighbour (KNN), and Support Vector Machine (SVM). The outcome exhibits the SVM with bigram in the Term Frequency Inverse Document Frequency (TF-IDF) procedure beat other mechanisms by 85%.

Chandra et al [7], have been provided methodology for sentiment analysis for the duration of the increase of COVID-19 patients in India by the deep learning long short-term memory (LSTM) [26] neural network mechanism. This developed skeleton take account of a state-of-the-art on LSTM, and BERT [25] vector embedding. Investigated the opinion about voiced over a few months in 2020, which coincide with India's peak in innovative cases. This classification employs multi label opinion segregation, which allows for the several emotions at the same time. Karim et al [17], have been carried out the sentiment analysis with a machine learning and rule-based mechanisms. In light of this measurement identified that from both of these methodologies machine learning mechanism is the best suitable for sentiment analysis. The rule based algorithms utilized the Senti word net, and Sentiment Vader, while the machine learning algorithm namely Naive Bayes utilized LDA.

Kaur et al [18], have been used Tweets data to investigate COVID-19, [1] new cases, deaths, recovered, and other keywords used to stream from

Twitter. This investigation developed new Hybrid Heterogeneous Support Vector Machine (H-SVM) to achieve opinion segregation and sort the scores as neutral, positive, and negative. Also match up to the developed system's performance with Support Vector Machine and Recurrent Neural Network (RNN) on the criteria namely recall, precision, accuracy, and f1 score. El Alaoui et al [12], have been gathered public opinion for analyzing immense social media data. Modern researches have utilized sentiment analysis and social media to follow public behavior in order to assist in key dealings. This study offered modifiable sentiment analysis mechanism, also capable to scan social media chats in real-time and retrieves individual opinions. This suggested mechanism creating a dictionary of words with valence, and then segregating tweets into diverse classes by the new additional elements that can optimize the valence degree of a review.

Shobana et al [24], have been suggested skip-gram manner for enhanced feature mining of contextual and semantic information. This advised process utilized the LSTM (long short-term memory) to realize difficulties [23] in the input of text and patterns. The adaptive particle Swarm Optimization (PSO) [27] procedure can optimizes weight factors to amplify the LSTM's performance. Experimentation on four kinds of dataset exposed that the advised APSO-LSTM process outperformed than classical VM, ANN, and LSTM processes in terms of accuracy.

3. RECENT STUDIES

[4] Several deep learning algorithms based on various embedding techniques were used in this work to contribute to the research topic of online customer evaluations. Our findings show that in a setting with fewer and more refined classes, all prediction models perform better, and that utilizing an augmented dataset improves prediction over the original dataset. When it came to context-free embeddings, Word2Vec outperformed FastText, although the differences were minor. Similarly, when compared to Bert and Albrecht, RoBERTa produced the best outcomes.

[9] The consequences of working out on the suggested model revealed that fastText defeats the SA-BLSTM and LSVM procedures, and that fastText is significantly suited to huge datasets on a server with negligible infrastructure. In comparison to the other two procedures, fastText delivered a more precise reaction in a shorter amount of time, with a rate of 90.71% in relation to LSVM and SA-BLSTM procedures.

[19] Several machine learning systems have been employed to categorize good, neutral, and negative reviews utilizing Bag of Words (BoW) and TD-IDF. The consequences show that integrating data balance with SMOTE improves the categorization accuracy. DT, SVM, and RF achieved 95% accuracy for SMOTE and BoW, and attained 95% accuracy with SMOTE and TF-IDF for SVM. The TextBlob dictionary is compared with the VADER dictionary and the SentiWordNet dictionary in terms of performance.

4. PROPOSED METHODOLOGY

Four different Amazon product categories have been covered in this suggested effort, namely Books, Mobile Products, Fashion Items, and Electronics Items. Based on the selected emotion features, lexical characteristics, and pragmatic features, the opinion mining task examined the collected Tweets and categorized the Tweets of product categories as positive or negative.

Typically, opinion mining is used to reveal the client's concealed opinion about an object or entity. In this study, an opinion mining system was developed to retrieve and mine opinion of the items from Tweets regarding Amazon product categories in order to categorize client opinions about the product as recommended or not recommended. NLP procedures and Python libraries have been used to preprocess the raw Tweets dataset, which has stored in HDFS. Tokenization, stop word deletion, stemming, and POS-tagging processes are all involved in the preprocessing phase. Lexical Features (Gram Words and Skip Gram Words), Sentiment Features (Opinion Words), and Pragmatic Features (Re-Tweets and Likes) are also chosen from the preprocessed Tweets dataset. The established valence approach was then used to recognize the chosen features modality words. Text Blob, Vader Sentiment, AFINN, and Senti Word Net are the four different valence calculating algorithms that make up this valence approach.

Finally, for each Tweet modality score estimated, pragmatic features criteria were established, and Tweets were classified as product recommended or not recommended by customers based on these rules.

This suggested work has been made possible by the Twitter API, Python libraries, and popular Natural Language Processing (NLP) procedures which allow streaming Tweets from Twitter. Tweets concerning product categories were classified and forecasted, then the experimental results were shown, and lastly the performance of the suggested mechanism was estimated using performance criteria. This recommended work has created and applied a Rule Based Classification and Prediction system for opinion mining, classification, and prediction tasks.

Algorithm of the Proposed Work

Input: Tweets

Output: Clients Recommended and Not Recommended Tweets

Step 1: Collect the Tweets

Step 2: Preprocess the Tweets

Input: Amazon product Tweets .csv file

Process: Tokenization, Deletion of Stop Words, Stemming, POS-Tagging

Output: Preprocessed Tweets .csv file

Step 3: Feature Selection and Opinion Valence Computation

1. Sentiment Features:
   (Valence of the Opinion Words)
2. Lexical Features:
   Valence of Gram Words: (Unigram, Bigram, and Trigram words)
   Valence of Skip Gram Words: (One-Skip-Gram, Two-Skip-Gram, and Tri-Skip-Gram words)
3. Pragmatic Features:
   (Re-Tweets and Likes)

Step 4: Rules Generation

Step 5: Classification Based on the Rules

Step 6: Performance Analysis

The previous research on opinion mining procedures has been allowed only Sentiment features, but the proposed work allowed Lexicon features and Pragmatic features. In this proposed technique three kinds of features have been collected and from the preprocessed tweets. According to the features sentiment values have been estimated for each work in tweets. And overall sentiment score for each tweet have been calculated. Then finally tweets have been classified and forecasted based on the generated rules.

For Lexicon features, sentiment features, sentiment value estimated. Then based on the pragmatic features rules have been generated for classification and prediction. These results can help to online buyers and sellers to know about the intended products. And can helps to make decision on shopping.

5. EXPERIMENTAL RESULTS AND DISCUSSION

This recommended classification system has been supported in the conducting tests by 6412 numbers of Tweets on Books, 9797 numbers of Tweets on Electronic Items, 53675 numbers of Tweets on Fashion Items, and 2725 numbers of Tweets on Mobile Products. On the streamed Tweets, sanitization methods have been applied, and opinion mining from the preprocessed Tweets was carried out. Opinionated Tweets have been separated using the produced criteria and the predicted Threshold value, and opinion forecasting was performed.

Table 1: Classified Opinionated Tweets Volume.

S.No	Types	Total Volume	Helpful Tweets	Not Helpful Tweets
1.	**Books**	6412	5265	1147
2.	**Electronic Items**	9797	9022	775
3.	**Fashion Items**	53675	50618	3057
4.	**Mobiles**	2725	2671	54

The Amazon product types Books, Mobile Products, Fashion Items, and Electronics Items are depicted in Table.1. Sanitized Tweets are represented by the field name Total Volume, while Helpful and Not Helpful Tweets are the segregated recommended and not recommended Tweets from the sanitized Tweets. 5265 records of cleaned Tweets about Books have been classified as Helpful Tweets, whereas 1147 records have been classified as Not Helpful Tweets. 9022 records of Tweets on Electronics Items have been classified as Helpful Tweets, while 775 records have been classified as Not Helpful Tweets, out of a total of 9797 records. Similarly, 50618 records have been classified as Helpful Tweets and 3057 records have been classified as Not Helpful Tweets from the cleaned 53675 records of Tweets on Fashion Items. As of the sanitized2725 data on Mobile Items, 2671 have been classified as Helpful Tweets, while 54 have been classified as Not Helpful Tweets.

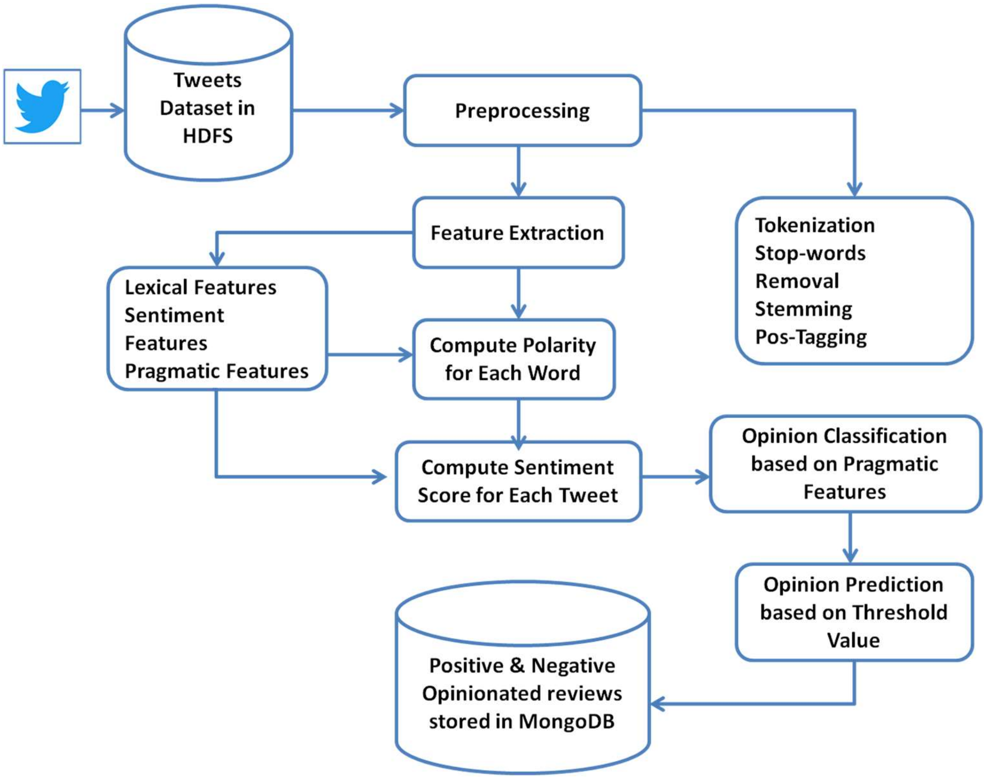
Tweets Dataset in HDFS
Preprocessing
Feature Extraction
Tokenization
Stop-words
Removal
Stemming
Pos-Tagging
Lexical Features
Sentiment
Features
Pragmatic Features
Compute Polarity for Each Word
Compute Sentiment Score for Each Tweet
Opinion Classification based on Pragmatic Features
Opinion Prediction based on Threshold Value
Positive & Negative Opinionated reviews stored in MongoDB

Figure 1: Architecture of the Proposed Work

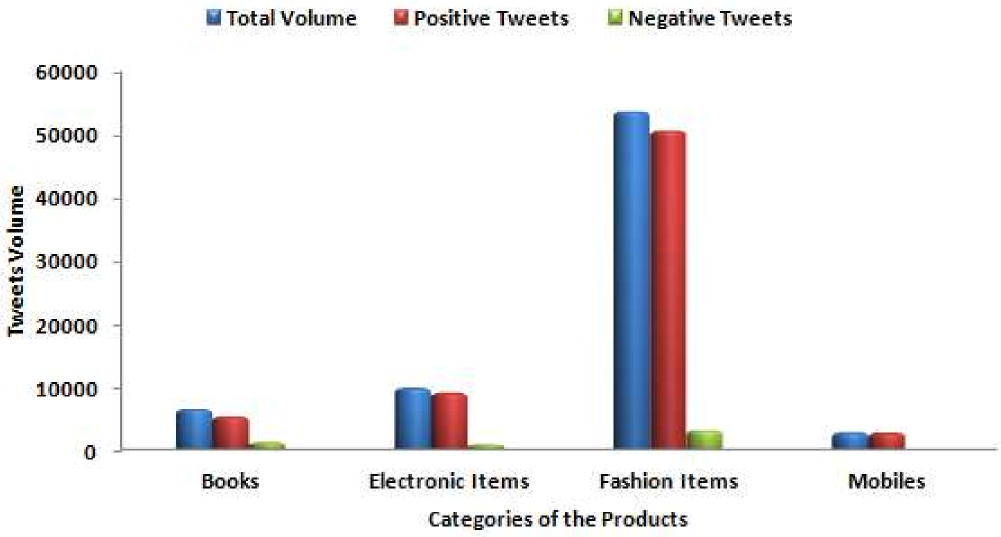

Figure 1: Classified Tweets using Proposed Technique

Table 1: Performance analysis Report on Proposed Classification Technique.

S.No	Categories	Accuracy	Precision	Recall	F1- Score
1.	**Books**	96.82%	100%	96.12%	98.02%
2.	**Electronic Items**	98.60%	100%	98.51%	99.25%
3.	**Fashion Items**	98.64%	100%	98.55%	99.27%
4.	**Mobiles**	99.41%	100%	99.40%	99.69%

Table.2 shows that the recommended Rule Based Classification and Prediction system has received superior ratings for performance evaluation criteria on Amazon product kinds such as Books, Mobile Products, Fashion Items, and Electronics Items. As a consequence, the recommended classification algorithm has achieved greater than 95% accuracy, recall, precision, and f1-score for all of the combined product Tweets dataset. These numerical representations demonstrate that the recommended and implemented system has been successfully implemented with reasonable values.

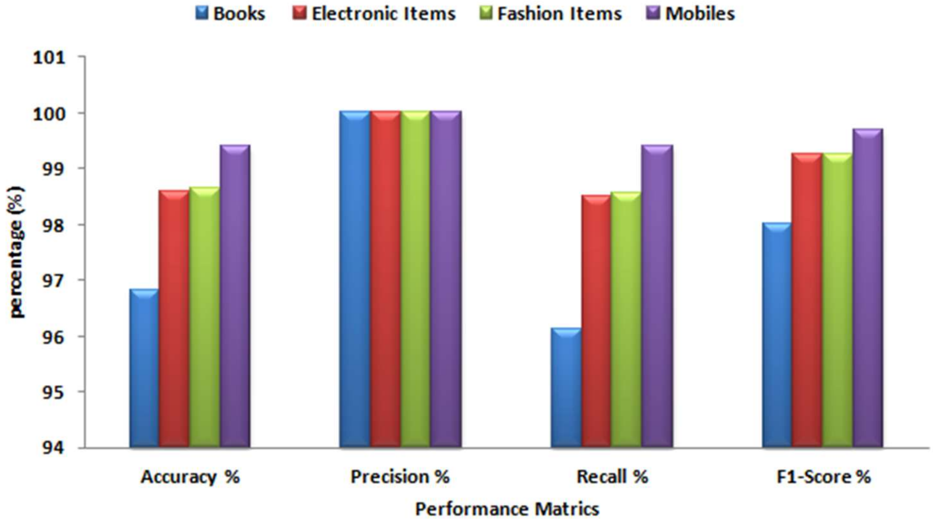

Figure.3. Performance Analysis Report for Proposed Work

Table.2 shows how the suggested opinion mining strategy performed in a performance analysis report on four different types of Amazon categories. Figure.3 shows how the proposed opinion mining approach performed in a performance analysis report on four different types of Amazon categories.

6. CONCLUSION

Buyer's opinions of Electronic Items, Fashion Items, Mobile Products, and Books on Amazon E-Commerce have been collected in the form of Tweets from Twitter in this suggested work. The proposed rule-based method has been used to assess variations in public opinion about various items on Twitter, categorizing them as positive or negative, and predicting buyer's view. Polarity scores have been generated for each opinionated phrase, and sentiment scores have been obtained for each Tweet, using the described sentiment score computation approach. Opinion has been separated and projected based on the expected Threshold value. This work suggests that the proposed work performed well and created greater accuracy based on the results of this proposed effort. As a consequence of this research, it appears that Amazon's Fashion Items, Mobile Products, and Books have a higher positive than negative opinion rate. These three products appear to have been well received and suggested by the buyers. However, the favorable opinion rate for electronic items is lower than the negative opinion rate. This indicates that Amazon Electronic Items are not well-liked or recommended by purchasers.

Limitation and Future Research Direction

In the collected tweets if the positive tweets volume is greater than the negative tweets volume then obviously the proposed opinion mining procedure can say that the product is recommended by the buyers. In others words, if the collected data imbalanced then the solution is not potential. In feature study imbalanced data will be take in account. That is to analyze the opinion mining positive and negative tweets will be taken in equal amount.

REFERENCES:

[1] Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., ... & Alaa, M. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert systems with applications*, *167*, 2021, 114155.

[2] Aljameel, S. S., Alabbad, D. A., Alzahrani, N. A., Alqarni, S. M., Alamoudi, F. A., Babili, L. M., ... & Alshamrani, F. M. A sentiment analysis approach to predict an individual’s awareness of the precautionary procedures to prevent COVID-19 outbreaks in Saudi Arabia. *International journal of environmental research and public health*, *18*(1), 2021, 218.

[3] Ayedh, A., Tan, G., Alwesabi, K., & Rajeh, H. The effect of preprocessing on arabic document categorization. *Algorithms*, *9*(2), 27. 2016.

[4] Balakrishnan, V., Shi, Z., Law, C. L., Lim, R., Teh, L. L., & Fan, Y. (2021). A deep learning approach in predicting products’ sentiment ratings: a comparative analysis. The Journal of Supercomputing, 1-21.

[5] Capuano, N., Caballé, S., Conesa, J., & Greco, A. Attention-based hierarchical recurrent neural networks for MOOC forum posts analysis. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 2021, 9977-9989.

[6] Ceron, A., Curini, L., & Iacus, S. M. iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences*, *367*, 2016, 105-124.

[7] Chandra, R., & Krishna, A. COVID-19 sentiment analysis via deep learning during the rise of novel cases. *Plos one*, *16*(8), e0255615, 2021.

[8] Chen, C., Sonnert, G., Sadler, P. M., Sasselov, D. D., Fredericks, C., & Malan, D. J. Going over the cliff: MOOC dropout behavior at chapter transition. *Distance Education*, *41*(1), 2020, 6-25.

[9] Chinnalagu, A., & Durairaj, A. K. (2021). Context-based sentiment analysis on customer reviews using machine learning linear models. PeerJ Computer Science, 7, e813.

[10] Dalipi, F., Zdravkova, K., & Ahlgren, F.. Sentiment Analysis of Students’ Feedback in MOOCs: A Systematic Literature Review. *Frontiers in Artificial Intelligence*, *4, 2021.*

[11] Dey, L., Chakraborty, S., Biswas, A., Bose, B., & Tiwari, S. Sentiment analysis of review datasets using naive bayes and k-nn classifier. *arXiv preprint arXiv:1610.09982, 2016.*

[12] El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A., & Kobi, A. A novel adaptable approach for sentiment analysis on big social data. *Journal of Big Data*, *5*(1),2018, 1-18.

[13] HaCohen-Kerner, Y., Miller, D., & Yigal, Y. The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, *15*(5), e0232525, 2020.

[14] HaCohen-Kerner, Y., Rosenfeld, A., Sabag, A., & Tzidkani, M. Topic-based classification through unigram unmasking. *Procedia Computer Science*, *126*, 2018, 69-76.

[15] Hassani, H., Beneki, C., Unger, S., Mazinani, M. T., & Yeganegi, M. R. Text mining in big data analytics. *Big Data and Cognitive Computing*, *4*(1),2020, 1.

[16] Jianqiang, Z., & Xiaolin, G. Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access*, *5*, 2017, 2870-2879.

[17] Karim, M., & Das, S. Sentiment analysis on textual reviews. In *IOP Conference Series: Materials Science and Engineering*, Vol. 396, No. 1, 2018, p. 012020.

[18] Kaur, H., Ahsaan, S. U., Alankar, B., & Chang, V. A proposed sentiment analysis deep learning algorithm for analyzing COVID-19 tweets. *Information Systems Frontiers*, *23*(6), 2021, 1417-1429.

[19] Machova, K., Mach, M., & Vasilko, M. (2021). Comparison of Machine Learning and Sentiment Analysis in Detection of Suspicious Online Reviewers on Different Type of Data. Sensors, 22(1), 155.

[20] Manguri, K. H., Ramadhan, R. N., & Amin, P. R. M. Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan Journal of Applied Research*, 2020, 54-65.

[21] Oliveira, D. J. S., Bermejo, P. H. D. S., & dos Santos, P. A Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology & Politics*, *14*(1), 2017, 34-45.

[22] Piryani, R., Madhavi, D., & Singh, V. K. Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management*, *53*(1), 2017, 122-150.

[23] Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2017). Challenges of sentiment analysis in social networks: an overview. *Sentiment analysis in social networks*, 2017, 1-11.

[24] Shobana, J., & Murali, M. An efficient sentiment analysis methodology based on long short-term memory networks. *Complex & Intelligent Systems*, *7*(5), 2021, 2485-2501.

[25] Su, Y., Xiang, H., Xie, H., Yu, Y., Dong, S., Yang, Z., & Zhao, N. Application of BERT to Enable Gene Classification Based on Clinical Evidence. *BioMed research international*, *2020*.

[26] Tiwari, A., Gupta, R., & Chandra, R. Delhi air quality prediction using LSTM deep learning models with a focus on COVID-19 lockdown. *arXiv preprint arXiv:2102.10551, 2021*.

[27] Wang, P., Zhao, J., Gao, Y., Sotelo, M. A., & Li, Z. Lane work-schedule of toll station based on queuing theory and PSO-LSTM model. *IEEE Access*, *8*, 2020, 84434-84443.