International Journal of Applied Sciences and Smart Technologies

Volume 4, Issue 1, pages 21-34

p-ISSN 2655-8564, e-ISSN 2685-9432

Impact of Online Education and Sentiment Analysis from Twitter Data using Topic Modeling Algorithms

Sulochana Devi[1, *], Chhaya Dhavale[1], Lalita Moharkar[2], Sushama Khanvilkar[3]

[1] *Department of Information Technology, Xavier Institute of Engineering, Mumbai, India*

[2] *Department of EXTC, Xavier Institute of Engineering, Mumbai, India*

[3] *Department of Computer Engineering, Xavier Institute of Engineering, Mumbai, India*

[*]*Corresponding Author: sulochana.d@xavier.ac.in*

(Received 01-05-2022; Revised 16-05-2022; Accepted 21-05-2022)

Abstract

During a pandemic, all industries suffer greatly, and every sector of the world suffers in some way, including the education sector. Internet expressions reflect users' feelings about a product or service. The polarity of information in source data toward a subject under investigation is determined by sentiment analysis processes. The goal of this study is to examine social media expressions about online teaching and learning, as online education will become a part of everyday life in the future. We collected data from Twitter using keywords related to online education and Google form from engineering undergraduate students for prototype implementation. This analysis will assist teachers, parents, and the student community in understanding the benefits and drawbacks of the education industry, allowing for further improvement in educational outcomes. We used aspect-based sentiment analysis and topic modeling to determine

This work is licensed under a Creative Commons Attribution 4.0 International License

Volume 4, Issue 1, pages 21-34
p-ISSN 2655-8564, e-ISSN 2685-9432

sentiment polarity and important topics for education sector stakeholders. To begin, we used TextBlob Python package to determine sentiment polarity, and Bag of Words, LDA and LSA model for discovering topics. After modeling topics from the collected data, topic Coherence is used to assess the degree of semantic similarity between high scoring words in the topic. The word cloud and LDAvis are used to visualize data. The experimental results are promising and it will assist education stakeholders in addressing the concerns that have been identified as social media expressions to work on.

Since the boom in science and technology, humans have been trying to invent machines that could reduce their efforts in day to day activities. In this paper, we develop a personal assistant robot that could pick up objects and return it to the user. The robot is controlled using an android application in mobile phones. The robot can listen to user's command and then respond in the best way possible. The user can command the robot to move to given location, capture images and pick objects. The robot is equipped with ultrasonic sensor and web camera that helps it to move to different location effectively. It is also equipped with sleds that play important role in object picking process. The robot uses a tiny YOLOv3 model which is rigorously trained on several images of the object. There are some possible improvements that can be achieved which could help this robot to be used in several other fields as well.

Keywords: coherence, education, LDA, LSA, pandemic, topic modeling, Twitter

1 Introduction

The global spread of COVID-19 altered humans daily routines of living, working, and operating in addition to their social connections. Like all other parts, the education sector grave implications related to students, instructors, and institutions across the

Volume 4, Issue 1, pages 21-34
p-ISSN 2655-8564, e-ISSN 2685-9432

world. Academic institutes were closed for formal offline face-to-face education to virtual transformation and the unprecedented upward thrust of on-line learning in the scenario of world COVID-19 lockdowns. Online learning, also known as e-learning, is learning via internet-enabled devices such as mobile phones, desktop computers, laptop computers, tablets, and so on. [1]. Within days, education was transitioned from face-to-face to online mode. All the stakeholders quickly adopted and continued with the new normal. This massive unplanned shift from traditional face-to-face learning to an online learning framework has altered the traditional ways in which educational institutions deliver knowledge to their students. Because of the abrupt transition, both teachers and students faced numerous challenges.

With video lectures and online exams, students were introduced to virtual textbooks and modules. Because of limited nonverbal communication, unstable internet access/technology, expensive equipment and devices, and a lack of technical knowledge, educators and learners faced numerous challenges in digital learning. Despite the challenges, online education provided many benefits such as continuity of studies, overall flexibility, increased information retention, extended reach of teaching, securing good grades, attendance, increased technical literacy, social interaction, accessibility, pace of learning, and no time and place restrictions.

Adoption of online learning will continue post-pandemic, and this shift will have an impact on the global education sector. A new hybrid educational model will emerge, with significant benefits. The integration of information technology in education will be accelerated, and online education will eventually become a required component of school education. Traditional offline learning and e-learning can coexist. Given the possibility of a hybrid model in which online learning will play an important role, there is a significant need to understand social media users' and learners' attitudes toward online learning in order to make online learning more effective.

This study examines students' perspectives on the positive and negative aspects of online learning, as well as the sentiments of Twitter users, who include students, teachers, and other stakeholders. During the pandemic, tweets from various education stakeholders such as teachers, students, parents, and other entities will cover the

Volume 4, Issue 1, pages 21-34
p-ISSN 2655-8564, e-ISSN 2685-9432

majority of the aspects of online learning. Data collected from students via Google forms include their perspectives on the abrupt shift from offline to online education, its impact on them, challenges encountered during the adaptation process, and their future expectations for online learning. The first tweets of education sector stakeholders were extracted using the snscrape scraper using various education-related keywords such as online education, e-learning, pandemic, and covid19. After preprocessing, sentiment analysis is done, as sentiments can be divided into positive, negative, and neutral forms, sentiment analysis processes identify the polarity of information in the source materials toward an entity. We applied the LDA model to discover topics from social media users' tweet data and students' opinions. The LDAvis tool was used for data visualization. LSA is applied to efficiently analyze the text and find the hidden topics of concerns related to education during a pandemic by understanding the context of the text. Bag of words technique is used to extract top words of concern for social media users. Same LDA, LSA and BoW techniques applied for data collected from students. [24] .

2 Research Methodology

Background work

So many researchers used twitter data for analysis. The reason could be that data can be extracted with API easily and it is one of the popular social media platforms. The collections of tweets contain useful information. Anyone can see the most recent expressions or complaints about any popular entity by combing through all of the tweets. The results of the analysis can be used to improve education performance, such as teaching and learning. Table 1 summarizes topic modeling algorithms and its applications used by various researchers. [1-3, 21,23].

Volume 4, Issue 1, pages 21-34

p-ISSN 2655-8564, e-ISSN 2685-9432

Table 1. Topic Modeling and Sentiment Analysis Review

Ref no of papers	Dataset Source	Approach	Aim/Objective	Remark
6	90,000 tweets from study area.	Naive-bayes classifier	Sentiment analysis of tweets on education during COVID-19	Topic modeling not done only sentiment analysis
7	1717 tweets from twitter	Web analytics approach	Find sentiment on educational posts	No ML
8	Online survey taken	The percentages were calculated based on the frequency of common student responses.	To know the effectiveness of online learning	Challenges and obstacles in online education at Pakistan students' perspectives
9	Data is collected with Google Forms	NLP techniques and Logistic regression classifier	Sentiment Analysis on COVID-19 Epidemic's Education	No topic modeling
10	Short Data from facebook	Topic modeling algorithm LSA, LDA, NMF, and PCA	Analysis of Topic modeling techniques	All topic modeling methods reviewed.
11	Facebook and Twitter	Topic modeling – LDA, LDAH, in the insurance domain. Sentiment analysis as well done.	Topic-aware sentiment analysis to improves, communication with customers and a better sense of the market	ML algorithms are used for text classification.
12	13,967 Tweet of Surabaya citizen	Topic modeling with LDA & LSA	Data requirement from community to support service policy,	Government made Surbaya media center.
13	1740 (Neural Information Processing Systems)	Topic modeling with LDA & LSA	Topic modeling to understand the various topics in fields of ML	Topic modeling used for detecting semantic structures in a set of research papers.

International Journal of Applied Sciences and Smart Technologies

Volume 4, Issue 1, pages 21-34

p-ISSN 2655-8564, e-ISSN 2685-9432

3 Materials and Methods

This section presents visualization and description of used datasets, description of sentiment analysis process, and the proposed methodology for performing topic modeling and sentiment analysis on the selected dataset.

Dataset Description

Two different types of data are used for this study. First dataset used for this study has been collected from Twitter through snscrape scraper using various keywords related to online education like 'onlineeducation', 'e-learning', 'pandemic', 'covid19', 'onlineclasses', 'educationincovid' etc. and contains 6000 records. Second dataset has been collected from 150 students using a google form to know their perspective about online learning, advantages, and disadvantages of online learning, difficulties and challenges faced during online learning. Table 2 presents sample tweets from the dataset with username, date and tweet content and Table 3 presents sample opinions from students' data.

Table 2. Tweet Sample from the collected dataset.

Username	Date	Tweet
Education blog	2020-10-17 14:43:06+00:00	#EDUCATION: As the #COVID19 #pandemic lingers, the impact on #highereducation is becoming clearer: #College closures, academic program terminations and institutional mergers are occurring at a pace seldom, if ever, seen before.
Paul Wusow	2020-10-16 20:23:04+00:00	The COVID-19 pandemic has changed education forever. This is how via @wef: https://t.co/owFvJYQZev #COVID19 #Pandemic #Education
Education blog	2020-10-15 15:16:23+00:00	#EDUCATION: The #COVID19 #pandemic is offering us an opportunity to rethink #accountability in education. https://t.co/jHTQ50Jrbt #schools #students #children #Teaching
panafrica nuk	2020-10-07 15:01:54+00:00	Teachers have had to move from a space in which they have years of experience to the unknown and challenging world of online, remote, correspondence and socially distanced teaching. Read more https://t.co/xhfuHRrxMh #Africa #AfricanCaribbean #COVID19 #Education #Pandemic

Volume 4, Issue 1, pages 21-34

p-ISSN 2655-8564, e-ISSN 2685-9432

Table 3. Sample Opinions from Students' dataset.

Sr. No.	Students' Opinion
1	Positive: No travel, comfortable environment, flexible Negative: No proper timeline, no proper instruments for practical for certain subjects.
2	Positive: Timings are flexible and we are doing it from the comfort of our homes, but the negative is the absence of social interaction and lack of routine'
3	Because of online education we get an opportunity to learn many new courses online and avoid the expenses of traveling.

After collecting both datasets, preprocessing is done to clean the dataset and remove unessential information. After that using TextBlob Python package polarity score of tweets is obtained. Tweets are categorized into three polarities: positive, negative, and neutral.

Methodology

This subsection explains different phases of the methodology followed and the approaches used in each phase. Aspect Based sentiment analysis and topic modeling techniques were applied on both datasets as shown in the given Figure 1.

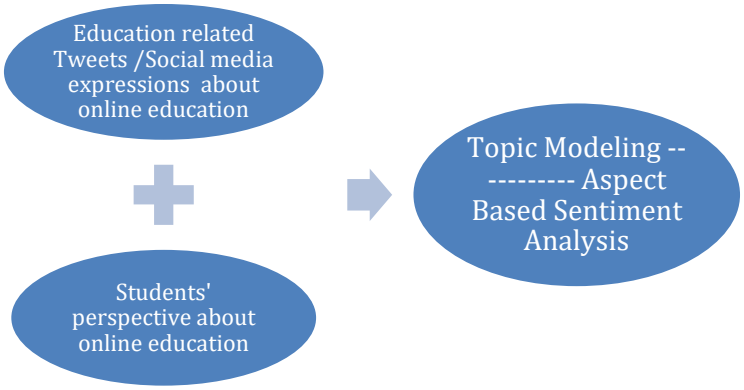


Figure 1. Proposed Work

The sequential workflow of the methodology applied on both datasets along with steps and methods is represented in the Figure 2. Workflow starts with the input phase of

Volume 4, Issue 1, pages 21-34

p-ISSN 2655-8564, e-ISSN 2685-9432

data collection from students and tweet scraping from Twitter using snscrap. In the next phase data is preprocessed to remove unnecessary and repeated words. After this phase sentiment analysis and topic modeling techniques are applied to obtain the required output.

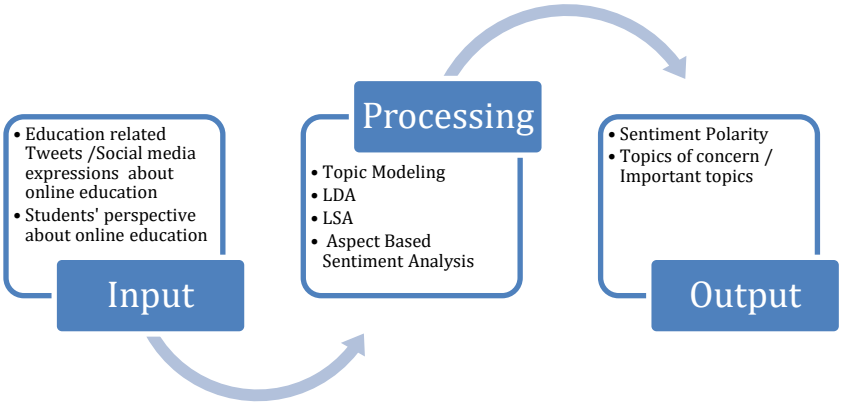


Figure 2. Processing Flow

Preprocessing of Data

Before starting data analysis process, data is preprocessed to remove non required information, this helps to increase efficiency of model and gives better accuracy. So, first step is to preprocess the data before starting encoding [3]. Python's NLP toolkit is used for preprocessing of data for this study. This is achieved by converting text into lower case, deleting URLS, removing hyperlinks and HTML tags, applying stemming, lemmatizing and finally removing stop words.

Lowercase conversion: changing the text to lowercase helps to decrease the complexity of the data as, 'data' and 'Data' are considered as different by machine learning models, so by changing all data into lower case, both words 'Data' and 'data' are considered as 'data'. Considering lower- and upper-case words as dissimilar words affects the training and classification process.

Elimination of URLs, punctuation marks, hyperlinks, HTML tags, and numbers: This type of data do not provide any additional meaning for learning models, so they do not contribute in enhancement of classification performance, also they and escalate the intricacy of feature set, so deleting them helps to bring down the feature space.

Volume 4, Issue 1, pages 21-34
p-ISSN 2655-8564, e-ISSN 2685-9432

Lemmatizing and Stemming: lemmatization and stemming is done to decrease inflectional forms and sometimes different forms of corelated words to a common base word form [4]. For example, ‘swims, ‘swimming’, and ‘swam’ are changed to the base word ‘swim’.

Elimination of Stop words: stop words do not provide any useful information for analysis. Stop words such as ‘yours’, ‘is, ‘the’, ‘a’, ‘am’ and ’an’ are removed [5].

4 Results and Discussion

TextBlob

TextBlob is a python library used for various NLP tasks such as sentiment analysis, part-of-speech tagging, paraphrase, noun phrase extraction, and sorting, etc. [14]. In our study, it is used for sentiment analyzing by providing polarity score between -1 and 1 for tweets. Tweets are assigned polarity based on polarity score, tweets having polarity score less than zero is considered as -ve, having score equal to zero is considered as neutral tweet, and having score greater that 0 will be considered as +ve tweet [15]. Table 4 presents polarity percentage of both datasets.

Table 4. Polarity Percentage

	Positive	**Negative**	**Neutral**
Dataset 1	38%	16%	46%
Dataset 2	63%	22%	15%

Feature Selection

Bag of Words (BoW) and TF-IDF are the most widely used methods for feature extraction. Bag of Words: BoW is a commonly used technique in NLP and information retrieval to extract features from preprocessed text or data [16].

BoW is used to count the appearance of a word in a text and forms a feature vector comprising the number of appearances of each unique word for text classification. The BoW is generally used to create the vocabulary of all unmatched words and train the learning models through their frequencies.

As an output of BoW, top words for dataset1 are: capacity, coffee, covid, education, experiences and for dataset 2 are: beginning, class, college, depressing, difficult etc.

Volume 4, Issue 1, pages 21-34
p-ISSN 2655-8564, e-ISSN 2685-9432

Term Frequency-Inverse Document Frequency: TF-IDF is used for feature extraction by extracting weighted features from text data. It gives the weight of each term in the corpus to enhance the performance of learning models [17].
TF-IDF score for search keywords such as education, covid, pandemic is **0.028766.**

Topic Modeling

For machine learning and natural language processing, topic modeling algorithm is used to scan large document, extract and phrase hidden patterns. Increased popularity of social media platforms makes them lucrative for researchers to extract ideas from here. Tweets contain unorganized short text topics, so it is required to uncover topics from tweet data through topic modeling.

In this paper, we have used LDA (Latent Dirichlet Allocation) and LSA (Latent Semantic Analysis) methods. LDA is an unsupervised generative probabilistic model of a corpus LDA gives topics using word probabilities. LAD has two parts, the words within documents, and probability of words is calculated related to a topic [23]. LDA is a well-known method for topic modeling. First introduced by David et al. in [22]. LSA is used to find out relation between documents and expressions. Performance of LSA is good in short sentence classification and it is demonstrated in various research works [20, 21]. Sample output obtained from LDA is given in Figure 3 & 4.

Topic Coherence is used to calculate the score of a single topic by calculating the degree of semantic similarity between high scoring words in the topic. These calculations help to discriminate topics that are semantically interpretable topics and topics that are artifacts of statistical inference. There are different coherence measures like c_v, c_p, c_uci, c_umass. We have used c_v and c_umass as given in the Table 5.

Table 5. Coherence Score

Coherence Score →	Using c_v	Using UMass
Dataset 1	0.41852	-6.95928
Dataset 2	0.33977	-2.48233

International Journal of Applied Sciences and Smart Technologies

Volume 4, Issue 1, pages 21-34

p-ISSN 2655-8564, e-ISSN 2685-9432

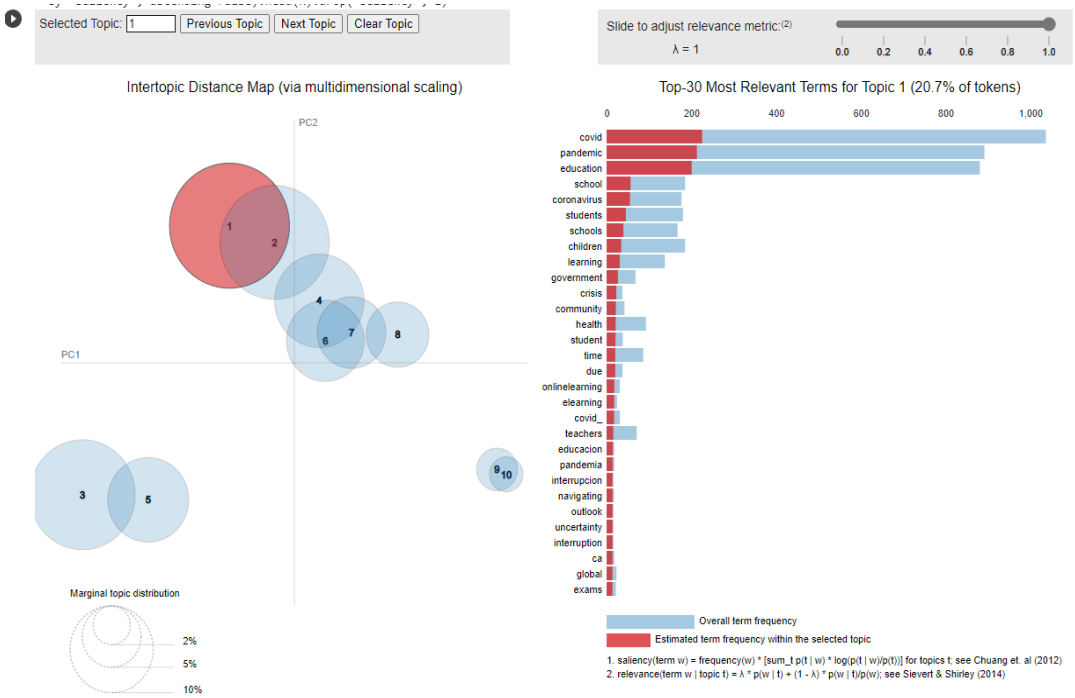


Figure 3. Top 30 Most Silent words in topic 1 twitter data

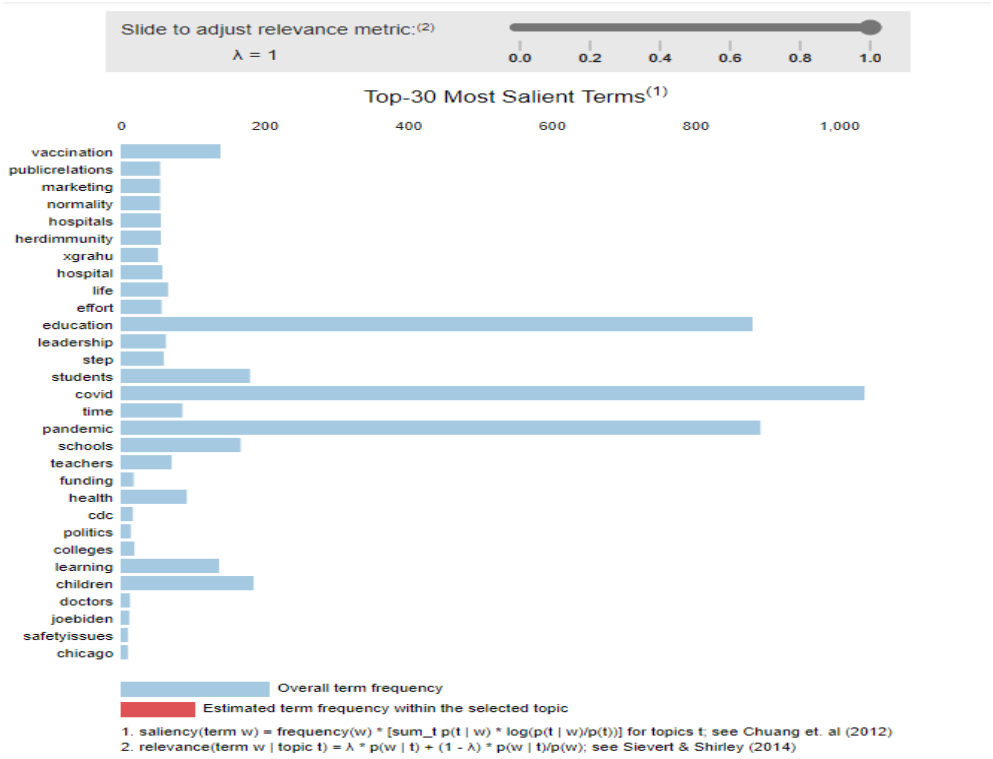


Figure 4. Top 30 Most Silent words in twitter data

Volume 4, Issue 1, pages 21-34

p-ISSN 2655-8564, e-ISSN 2685-9432

5 Conclusion

This study examines a topic related to online education during the Corona period. Polarity is calculated and analyzed as positive, negative, and neutral for the same dataset. Two datasets are created: one for Twitter data and its analysis, and the other data collected from students to understand the impact of online education on the student community. The LDA and LSA algorithms have been used successfully for topic modeling. LSA is a method for forming semantic generalizations from textual sections that uses Singular value decomposition (SVD), whereas LDA is an unsupervised machine learning algorithm. Model-generated topics are not always easy to interpret. Topic coherence calculations are thus used to differentiate between good and bad topics. The outcomes / topics can then be used to overcome challenges in the education industry, as well as to consider online education as an opportunity for the needy. We discovered clear polarity/understanding in the student dataset and mixed expressions in the social media community.

Acknowledgements

Authors are highly grateful to twitter for providing a platform for users to express their views, and thankful to Xavier Institute of Engineering, Mahim Mumbai.

References

[1] Zhu, X., & Liu, J. Education in and After Covid-19: Immediate Responses and Long-Term Visions, *Postdigital Science and Education*, **2**(3), 695–699, 2020.

[2] Mujahid M, Lee E, Rustam F, Washington PB, Ullah S, Reshi AA, Ashraf I. Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19, *Applied Sciences*, **11**(18), 8438. 2021.

[3] Reddy, A.; Vasundhara, D.; Subhash, P. Sentiment Research on Twitter Data, *Int. J. Recent Technol. Eng.*, **8**, 1068–1070, 2019.

[4] Jivani, A, A Comparative Study of Stemming Algorithms, *Int. J. Comp. Tech. Appl.* **2**, 1930–1938, 2011.

Volume 4, Issue 1, pages 21-34

p-ISSN 2655-8564, e-ISSN 2685-9432

[5] Armstrong, P., Bloom's Taxonomy. *Vanderbilt University Center for Teaching*. 2019.

[6] Cheeti, S. S. Twitter based Sentiment Analysis of Impact of COVID-19 on Education Globaly, *International Journal of Artificial Intelligence and Applications (IJAIA)*, **12**(3), 2021.

[7] Floradel S. Relucio and Thelma D. Palaoag. Sentiment analysis on educational posts from social media, *In Proceedings of the 9th International Conference on E-Education, E-Business, E-Management and E-Learning (IC4E '18). Association for Computing Machinery*, New York, NY, USA, 99–102, 2018.

[8] Adnan M, Anwar K., Online learning amid the COVID-19 pandemic: Students' perspectives, *Journal of Pedagogical Sociology and Psychology*. **2**(1), 45-51. 2020.

[9] Sanjok Lohar, The Impact Ofcovid-19 Pandemic On Education System, *International Journal of Emerging Technologies and Innovative Research*, **8**(4), 428-430, April 2021.

[10] Albalawi, R., Yeap, T. H., & Benyoucef, M., Using Topic Modeling Methods For Short-Text Data: A Comparative Analysis, *Frontiers in artificial intelligence*, **3**, 42, 2020.

[11] Albalawi, R., Yeap, T. H., & Benyoucef, M, Using Topic Modeling Methods For Short-Text Data: A Comparative Analysis. *Frontiers in Artificial Intelligence*, **3**, 42, 2020.

[12] Qomariyah, S., Iriawan, N., & Fithriasari, K. Topic modeling twitter data using latent dirichlet allocation and latent semantic analysis, *In AIP conference proceedings* **2194**(1), 020093. AIP Publishing LLC, December 2019.

[13] Slimane Bellaouar, Mohammed Mounsif Bellaouar, and Issam Eddine Ghada. Topic Modeling: Comparison of LSA and LDA on Scientific Publications. *In 2021 4th International Conference on Data Storage and Data Engineering DSDE Association for Computing Machinery*, New York, NY, USA, 59–64. 2021.

[14] Loria, S. textblob Documentation. Release 0.15. **2**, 269, 2018.

[15] Sohangir, S., Petty, N., & Wang, D, Financial Sentiment Lexicon Analysis, *IEEE 12th International Conference on Semantic Computing (ICSC)*, 286-289, 2018.

[16] Eshan, S.C., & Hasan, M.S., An application of machine learning to detect abusive Bengali text, *20th International Conference of Computer and Information Technology (ICCIT)*, 1-6, 2017

[17] Zhang,W.; Yoshida, T.; Tang, X. A comparative study of TF* IDF, LSI and multi-words for text classification, *Expert Syst. Appl.* **38**, 2758–2765, 2011.

[18] Robertson, S., *Understanding inverse document frequency: on theoretical arguments for IDF*, Journal of Documentation, **60**(5), 503-520, 2004.

[19] George, M., Soundarabai, P.B., & Krishnamurthi, K., Impact Of Topic Modelling Methods And Text Classification Techniques In Text Mining: A Survey, 2017.

[20] Salloum, S.A., Al-Emran, M., Monem, A.A., Shaalan, K, Using Text Mining Techniques for Extracting Information from Research Articles. *In: Shaalan, K., Hassanien, A., Tolba, F. (eds) Intelligent Natural Language Processing: Trends and Applications. Studies in Computational Intelligence*, **740**. Springer, Cham. 2018.

[21] Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by latent semantic analysis. *J. Am. Soc. Inf. Sci.* **41**, 391–407, 1990.

[22] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022, 2003

[23] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools Appl*, **78**, 15169–15211. 2019.

[24] Mujahid M, Lee E, Rustam F, Washington PB, Ullah S, Reshi AA, Ashraf I. Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19, *Applied Sciences*. **11**(18), 8438, 2021.