



Article

Detection of Students' Problems in Distance Education Using Topic Modeling and Machine Learning

Huda Alhazmi

Department of Computer Science, Umm Al-Qura University, Makkah 24236, Saudi Arabia; hnhazmi@uqu.edu.sa

Abstract: Following the rapid spread of COVID-19 to all the world, most countries decided to temporarily close their educational institutions. Consequently, distance education opportunities have been created for education continuity. The abrupt change presented educational challenges and issues. The aim of this study is to investigate the content of Twitter posts to detect the arising topics regarding the challenges of distance education. We focus on students in Saudi Arabia to identify the problems they faced in their distance education experience. We developed a workflow that integrates unsupervised and supervised machine learning techniques in two phases. An unsupervised topic modeling algorithm was applied on a subset of tweets to detect underlying latent themes related to distance education issues. Then, a multi-class supervised machine learning classification technique was carried out in two levels to classify the tweets under discussion to categories and further to sub-categories. We found that 76,737 tweets revealed five underlying themes: educational issues, social issues, technological issues, health issues, and attitude and ethical issues. This study presents an automated methodology that identifies underlying themes in Twitter content with a minimum human involvement. The results of this work suggest that the proposed model could be utilized for collecting and analyzing social media data to provide insights into students' educational experience.

Keywords: distance education; topic modeling; biterm topic model (BTM); machine learning; support vector machine; Twitter



Citation: Alhazmi, H. Detection of Students' Problems in Distance Education using Topic Modeling and Machine Learning. *Future Internet* **2022**, *14*, 170. <https://doi.org/10.3390/fi14060170>

Academic Editor: Filipe Portela

Received: 28 April 2022

Accepted: 29 May 2022

Published: 31 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The coronavirus pandemic has wreaked havoc on every aspect of our modern society. The world was challenged on an unprecedented scale. In response, most governments around the world authorized containment measures including social distancing and the temporary suspension of schools and educational institutions. Educational pedagogies were moved to online education. In some countries, distance education was adopted to maintain the educational activities [1–3]. As a result of the vital need to replace face-to-face learning, educational institutions had to adopt a digital approach where teaching is delivered on digital platforms [4–8].

Contrary to general belief, distance education is not a new phenomenon. The emergence of distance education dates back to the 18th century. Countries such as the USA, England, and Australia have provided distance education in the form of correspondence study to reach learners who are geographically dispersed [9]. With the rapid development in technology, several models in distance education have developed such as online learning in which students use Internet and associated technologies to learn from their homes [10] and blended learning that combines online and in-person learning [11]. All these models use technology in learning but adapt different learning processes. While there has been an increase in the use of online distance learning in various countries, some challenges have been posed for students, parents, teachers, and institutions due to the abrupt transition.

As the use of online distance learning continues to increase, the need for quality data to detect underlying trends in order to optimize the efforts of improving the educational experience is essential. Currently, existing research largely relies on surveys through

interviews or questionnaires [8,10,12]. The collected data from these surveys are critical in identifying the themes regarding to the perspectives about distance learning experience. However, their reliability relied on the respondents. Furthermore, generating the result from these data commonly takes time; meanwhile, the observed trends might be changed by the time the result is reported. Hence, alternative methods are needed, including analyzing the data of social media platforms such as Twitter, Facebook, and YouTube. Such platforms offer open sources to students to share their experience and seek social support. This vast amount of data provides educational researchers with implicit knowledge and new perspectives to better understand the students' experiences that inform of the improvement of education quality [13].

Despite the challenges associated with analyzing this massive amount of data to obtain valuable insight, the research collaborations and improvement increased. In response, this study presents a new approach using Twitter data in order to gain insights of distance education issues. Our focus is on Arabic Twitter content posted by users related to distance education in Saudi Arabia at all educational levels. The aim of this study is to detect and classify the potential topics in social media content related to distance education to identify the major issues students face through their distance education experience. To attain this goal, the objectives of this study are as follows:

- Collect Arabic tweets regarding distance education in Saudi Arabia and perform filtering and preprocessing process.
- Apply and train biterm topic model (BTM), an unsupervised machine learning approach on the cleaned tweets to identify the underlying topics and cluster them into themes.
- Annotate the resulting themes into major categories and sub-categories that represent the issues of the distance education.
- Use multi-class SVM classifier model to classify the entire dataset into the generated categories from the previous step to identify the major issues.

In this study, we proposed a new methodology that integrates clustering and classification machine learning techniques to explore the textual Twitter data shown in Figure 1. Hence, the study differs from prior work in implementing the automatic filtering of a large dataset of tweets and detecting topics that are contained in the data in order to inform educational decision makers about the issues and problems that the students face in their learning experience.

The rest of the paper is organized as follows: Section 2 introduces related work. Section 3 contains the illustration and implementation of the proposed method. Section 4 shows the results. Section 5 presents the discussion and limitation followed by the conclusion in Section 6.

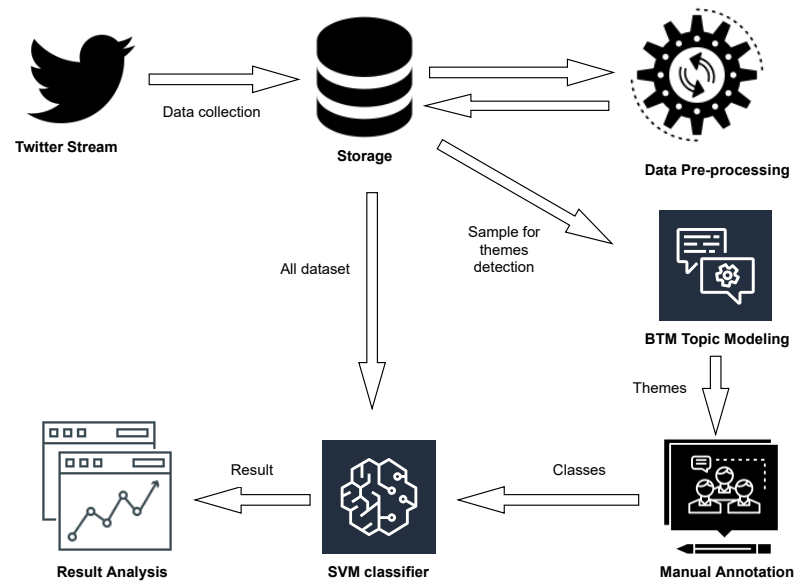


Figure 1. The workflow of the proposed method.

2. Related Work

Extracting the topics that reflect students' concerns according to their online conversations has become an important source to enhance distance education and develop a new learning styles. Consequently, many studies have been conducted extensively. In this section, we will focus on studies related to the topic detection methods and educational data analysis.

The widely used models to uncover the hidden topics in the text are the probabilistic semantic-based indexing model (PLSI) [14] and latent dirichlet allocation (LDA) [15]. PLSI underlines the topic based on the semantic indexing wherein low-dimensional representation has been utilized to evaluate the similarities. The LDA model is an unsupervised model that has been used to discover and document subject knowledge. For online communities, several models were proposed based on the LDA model to mine the online content and identify the topics. Nagori et al. developed a personalized recommendation system for e-learning based on the content of the text. They utilized the topic model by introducing the similarity measures [16]. Zhong et al. [17] designed a model for the quality of online students' comments; they constructed a set of topic features that strongly related to the quality comments. Applying such models directly on short texts will yield to data sparsity problem [18]. Yan et al. [18] proposed a model to detect topics in short textual data called Biterm Topic Model (BTM); their model generates the term co-occurrence in the whole corpus to improve the topic learning.

Distance education is a system of learning and teaching. Analyzing and mining the data generated in an online education setting is an effective method to mine and understand students' learning environments. Data-driven approaches have been used to analyze structured data including student performance, activity, or course discussion forums [19]. In fact, the exchange of the information about online education has been increased using online platforms, and various studies have been put forward to enhance the quality of data mining in distance learning. Kagklis et al. [20] utilized text mining techniques to improve the educational process; they used regression techniques to predict the performance of the students. Rooyen investigated students' opinions in the accounting department at the University of South Africa about integrating social media apps in the learning process to improve their education. They found that students have positive opinions toward using social media for academic support [21]. The authors in [22] proposed an intellectual distance education system to exchange information of artificial immune systems. The system allows learners to experience a real-time training to gain necessary skills.

The opinion mining system has been widely used to measure public opinion about any topic or event. Many studies have been carried out to investigate the opinion regarding distance education. Kumar et al. [23] developed a multi-aspect-based system to explore students' opinion about distance education in India. They applied sentiment classification on online reviews using Naive Bayes classifier. A recent study by Aljabri et al. [24] employed a sentiment analysis approach to investigate people's opinions about distance learning in Saudi Arabia. They utilized different classification techniques including support vector machine, Naive Bayes, logistic regression, k-nearest neighbor, and XGBoost, and the best accuracy was achieved in logistic regression with a unigram feature. Their finding demonstrates that there was a general positive opinion at the school level and negative opinion at the higher education level. A more recently study by Alqurashi [25] applied several classical machine-learning models to explore public's opinion in KSA regarding distance education. The author classified the tweet into two classes (Infavor, Against), which represent negative and positive opinion. The result of the study showed positive opinion related to distance education.

Regarding the challenges, research has identified challenges concerning distance education. A study by Abuhammad [26] applied the qualitative method to investigate the challenges of distance learning with regard to parents' perceptions during the coronavirus outbreak in Jordan. The analysis was applied on Facebook groups with a total of 248 posts. The finding revealed four themes, naming personal barriers, technical barriers, logistical barriers, and financial barriers. A recent study by Almendingen et al. [27] developed a methodology to assess the Norwegian faculty and students' experiences for shifting to distance learning. They utilized a mixed methods approach that combined quantitative and qualitative analysis. They found that students adapted quickly to the new learning style but also experience difficulties such as lack of social interaction, limited participation, and stressful situations. A more recent study by Segbenya et al. [28] examined students' perspectives about distance education. They employed a descriptive statistics method using a survey to analyze students' learning challenges. Their finding identified two main challenges, including unreliable internet and power and lack of collaboration and motivation.

3. Materials and Methods

This study was conducted in three phases: (1) data collection using public Twitter API followed by subsequent preprocessing steps to filter and clean the tweets; (2) topic modeling using an unsupervised machine learning approach to identify the underlying themes; and (3) classification using a supervised machine learning approach to classify the whole dataset into categories and sub-categories. The structure of the method is shown in Figure 2. Detailed descriptions of these phases are in the following subsections.

3.1. Data Collection and Preprocessing

The data were collected using Twitter API; we first collected Arabic tweets containing specific keywords including (التعليم عن بعد), (الدراسة عن بعد), and (الدراسة اون لاين). The translation of these keywords is "distance education", "distance learning", "online education", and "online learning". We streamed tweets for about three months from 1st July 2020 to 29 September 2020. A total of 2,811,327 tweets were collected. We further filtered this corpus for tweets belonging to Saudi Arabia, resulting in 901,938. Additionally, repeated tweets and tweets not related to user content such as tweets contain news or commercial information were removed, resulting in 207,488, as shown in Table 1.

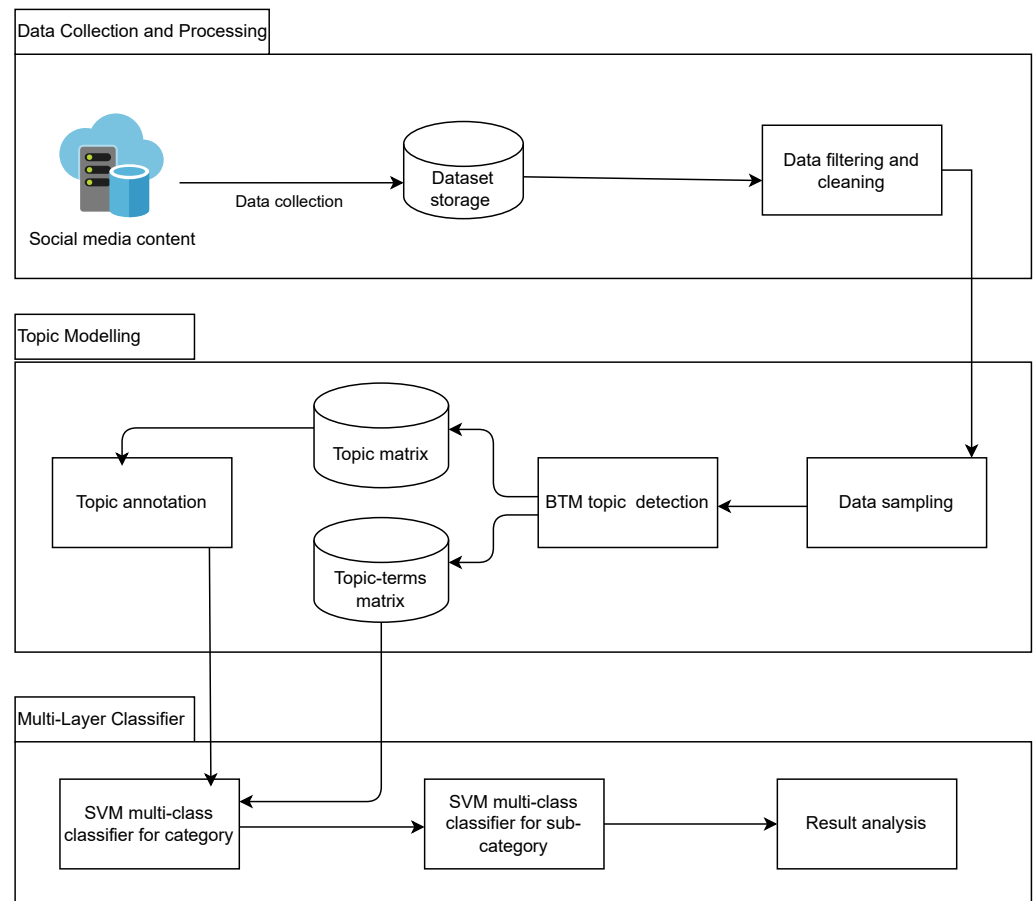
Twitter posts include special symbols to convey certain meaning; also, the collected tweets may contain lots of noisy and uninformative data. Therefore, the tweets were cleaned by several standard data preprocessing steps.

- Convert text files to UTF-8 encoding.
- Remove punctuation marks, non-letters, and diacritics.

- All tweet's features include hashtags, mentions (@user), URLs, and retweet symbol (RT) were removed.
- Replace the repeated letters such as (تعبااان) with two letters (تعبان) to identify the word.
- All Arabic stop words such as (اين , عن) were removed.

Table 1. Number of collected tweets.

Month	Number of Collected Tweets	Number of Tweets after Cleaning and Filtering
July	959,664	53,563
August	1,089,240	108,924
September	762,423	45,001
Total	2,811,327	207,488

**Figure 2.** The structure of the method.

3.2. Topic Modeling

Topic modeling is a probability generation for text content to discover the hidden themes and their vocabulary. The focus of this study is to identify what were the major issues that students encounter through their distance education experience in Saudi Arabia. We used the biterm topic model (BTM), an unsupervised machine learning approach [18] as a core for discovering the themes in our corpus, where each tweet is treated as a document. BTM is a mathematical method developed to work with a sort text. A graphical representation of this model is shown in Figure 3 [18]. BTM works by generating the co-occurrence of all word-pairs which are called biterms within each text across the corpus. The generative process of BTM is presented in Equation (1).

$$P(B) = \prod_{i,j} \sum_z \theta_z \phi_{i|z} \phi_{j|z} \quad (1)$$

An implementation of BTM was provided by bitermplus package [29]. In the learning step, the tweets corpus and the number of topics (k) are fed to the BTM. Then, the model generates the topics themes as word groupings and their discrete probability distribution. The highest probabilities were assigned to the words that are the most representative of the theme. We implemented the algorithm by using bitermplus package [29] in Python language version 3.7.10. We developed the themes in two stages; the first involved the implementation steps of the BTM approach, and the second involved the manual annotation to identify the themes that related to the issues of the distance education. In the following, we will explain these stages in detail.

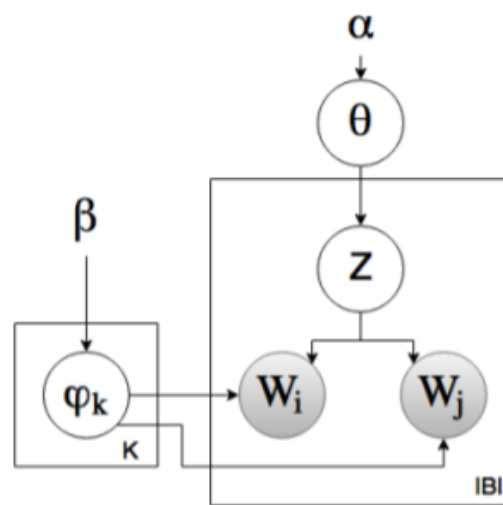


Figure 3. The graphical representation of the BTM model.

3.2.1. BTM Implementation

The following steps are performed to train the BTM and extract the themes.

- **Selecting the training sample:** In this step, random sample was extracted from the corpus to train and fit the BTM model. All empty tweets or with less than three words generated from the preprocessing step were filtered. Then, the resulted 5000 tweets were used for training.
- **Estimating the number of topics (k):** Identifying k is very important step because it can affect the produced topics. A large k could lead to many clusters that can not be interpreted, while a small k could lead to a general cluster in term of semantic contexts. To define the optimal k, different qualitative or quantitative approaches have been used. We used a coherence score to measure the quality of learned topics; a higher score means high correlation of the topics in the cluster. The UMmass coherence score was used to calculate the scores as shown in Equation (2) [30]. $D(w_i, w_j)$ represents the documents count in which words w_i and w_j co-occur, where 1 is added to this term to avoid calculating the logarithm of zero when the two words never appear together. This step was repeated with different values of k, ($k = 5, 7, 10, 15$) to ensure that the filtered topics are highly relevant to the issues of interest. We found the best value of k was 5.

$$SCORE_{UMass}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)} \quad (2)$$

- Training the BTM model: In the training phase, we trained and fitted BTM with $k = 5$ as estimated in the previous step. Then, we applied the inference phase; in this phase, the algorithm determines how each tweet correlates to the themes discovered in the learning phase. As a result, the themes were produced with five tweets in each.

A sample of the result of this stage is demonstrated in Table 2. Notably, there is an overlap between the themes, which might be because the tweets represent very connected issues in distance education.

Table 2. Sample of the themes.

Theme	Example
Theme 1	الكوزات الغش بفلس اشتقنا المدرسه استاذتنا ساحه مدرستنا شاء الل تعود الايام شكل جميل وتذهب قسوه الظروف التعليم المدارس جميل شوق التعليم الالكتروني سيء بمعنى الكلمه اقتدار الحصص المسجله التواصل الصفوف
Theme 2	قسم الطلاب محاضره يفتحون البرنامج ينامون الطلاب يعاني عدم وصول الشبكه سبب التعليم بعد طلبه الجامعات الدوام الحضورى الظروف صعبه التعليم بعد سيء شبكه ظل ظروف التعليم بعد الحاجه الماسه خدمه النت تعطلها تعطل طالب اتفقنا الدراسه بعد اسوء كورونا نفسها
Theme 3	طلبه الجامعات الدوام الحضورى الظروف صعبه التعليم بعد سيء شبكه التعليم التواصل السيء البعيد الدكاتره المشرفين جدولي يحتظر الوقت تفكرني حاضره مركزه نفس الوقت اسوي بسبوسه الدراسه بعد ساتره حاجات كثيره نجاح الطالب التعليم بعد يثبت ذكاء الام فضلها حل دروسه امتحاناته يلعب
Theme 4	تعبت نفسيا الدراسه بعد افهم شي استوعب كرهت الدراسه الل ياخذ التعليم بعد وربي شي يقهر التعليم بعد يقارن الحضورى باذن الل اخر سنه يطبق فجوه كبيره التعل احد الاسباب يكرهك تشارك المايك خربان
Theme 5	اقتدار الحصص المسجله التواصل الصفوف فاتحين البرنامج فاتحين بونجي اشتقنا المدرسه استاذتنا ساحه مدرستنا شاء الل تعود الايام شكل جميل تذهب قسوه الظروف الاستاذ تشرح اسوي هوت شوكلت التعليم المدارس جميل شوق

3.2.2. Topics Annotation

The resulted themes from the BTM model were annotated manually by two annotators: the author and one of the colleague as a native Arabic speaker. Each annotator developed a set of categories independently. Then, these categories were narrowed down into 11 notable themes. In addition, these themes were considered as sub-categories and grouped in general categories. The agreement between the two annotators achieved a high inter-rate with $\kappa = 0.91$. These sub-categories have been identified to cover the issues that students experienced through their educational process. Each sub-category belongs to one of the five general categories: education issues, social issues, health issues, technology issues, and attitude and ethical issues. Table 3 lists the top 20 terms with highest probabilities in all selected topics.

Table 3. The top 20 terms for each of the general topics.

Education Issues		Social Issues		Technology Issues		Health Issue		Attitude and Ethical Issue	
الدراسة	Studying	اشتقنا	We miss	النت	Net	نفسيا	Psychologically	غش	Cheating
بعد	Distance	الوقت	Time	الشبكة	Internet	تعبت	Get tired	نوم	Sleep
الشخصي	Personal	التعليم	Education	وصول	Access	كرهت	Hate	سهر	Stay up
اهميه	Important	الالكتروني	Electronic	علق	Hang	صداع	Headache	لعبه	Game
الحضور	Attending	يفصل	Separates	خدمه	Service	عصبي	Nervous	تشرح	Explain
التواصل	Communication	افتراضي	Virtual	تعطل	Broken down	الغياب	Absence	نلعب	Play
السيء	Bad	تمود	Return	برج	Tower	شيء	Thing	الاستاذ	Teachers
اولياء	Parents	المدارس	Schools	المايك	Mic	سيء	Bad	التركيز	Focus
البعيد	Remote	شوق	Longing	اللابتوب	Laptop	استوعب	Realize	حضور	Attendance
الدكاتره	Professor	التواصل	Communication	الجهاز	Device	نوم	Sleep	ذكاء	Intelligence
مستوعب	Understand	الاصدقاء	Friends	خرابان	Broken	مراعاة	Caring	اولياء	Parents
التركيز	Focus	الواقع	Reality	ظروف	Circumstances	يتحمل	Endure	التعليم	Education
سرحان	Stray	جميل	Nice	بعد	Distance	ضغط	Stress	بعد	Distance
الحصص	Classes	الايام	Days	التعليم	Education	يصير	Occur	اجازه	Vacation
صعبه	Difficult	توتر	Anxiety	الصوت	Sound	صعب	Difficult	امتحانات	Exams
فاشل	Failure	الظروف	Circumstances	يقطع	Disconnect	صبر	Patience	حل	Solution
الاختبارات	Tests	اساتذتنا	Our teachers	واضح	Clear	يقهر	Vanquish	اختبار	Test
الالكتروني	Electronic	الكلمه	Word	ضعيف	Weak	كميه	Amount	الوقت	Time
المدرس	Teacher	جو	Atmosphere	صعوبه	Difficulty	الدراسة	Studying	تدخل	Interfere
التعليم	Education	الدراسة	Studying	يسجل	Recording	بعد	Distance	الحصه	Class

3.3. Multi-Layer Classification Model

Many approaches in the literature have been developed to address a multi-class classification problem. However, the most common solution is to decompose the multi-class classification problem into a set of binary classification problems. One of these methods is called one-versus-all [31]; the basic concept of this method is conducting a binary classifier for each label with assuming independence among the labels. Based on this mechanism, we performed a multi-class learning by using standard support vector machines (SVMs). If we have k classes, then SVM classifiers are constructed for each class. Every classifier is trained to distinguish the samples of one class from the samples of all the other classes. To predict a new sample, the sample is classified by all k classifiers, and then, the class with the maximum decision value is chosen.

Our classification approach includes two layers of a multi-class SVM classifier. The topics that were produced from the annotation stage in the second phase are used as classes for classification of the tweets in the entire dataset. The general set of classes is (education issues, social issues, technology issues, health issues, and attitudes and ethical issues), which represents the general categories. The other set of classes is (professor–student interaction, lack of focus, student–student interaction, lack of time, lack of social interaction, poor internet connectivity, lack of tech equipment, overuse of digital technologies, stress and anxiety, lack of motivation, and cheating) that represents the sub-categories. The first classifier is trained to predict if the tweet belongs to one of the basic categories. The second classifier is used to predict if the tweet belongs to sub-categories. We implemented the two classification tasks in a multi-layer classification model. In the upper layer, the multi-class classifier for the categories is applied. Then, the predicted category will be passed as a feature to the lower layer to classify the tweet as a sub-category. To avoid the misclassification that could occur on the upper level, all sub-category classifiers are applied in the lower level.

Features extraction is an essential step in the classification process in order to represent the tweet text as numerical vectors. For features extraction, we extracted two features as follows:

- The top 20 words with the highest probability of occurrence that are extracted from BTM for each category are used to build a set of features that is used for classification in the first layer.
- N-gram is a set of consecutive order of words based on the value of N. We extracted the bigram where $N = 2$ is to be used for classification in the second layer.

The experiment is performed using Python language version 3.7.10; we used scikit-learn library [32] to implement the multi-class SVM classifiers. The random sample of tweets that was used to train BTM algorithm was also used to train and fit the multi-class SVM classifier. We first partitioned the data into 70% training set and 30% test set to train and fit the classifiers. Then, we applied the classifiers on the rest of the data.

4. Results

In this section, we first analyze the statical implication of the count regarding the detected categories. Afterwards, precision, recall and F1 measures are used to evaluate the performance of the classiiier model.

4.1. Statistical Analysis

Figure 4 shows the number of tweets in each category of a 5000 tweets sample that is used for developing these categories. Each category represents one of the students' problems in distance education. We consider each tweet to fall only in one category, as we note that the largest category is education issues and the smallest category is attitudes and ethical issues, whereas the large number of tweets fall in 'Others'. Multi-class SVM is used to classify students' problems, 37% of 207,488 cleaned tweets were classified as general and sub issues. Figure 5 and Table 4 show the counts and percentages of general and sub issues that were detected. The category with the most frequent mentions is education issues with 23,976 tweets, which is followed by social issues with 17,701. This implies that these issues are considered highly important from the students' point of view. Surprisingly, technology issues is the third most frequent category, which is not what we expected to find. We also notice that a quite small percentage of tweets reflects issues related to attitudes and ethical category.

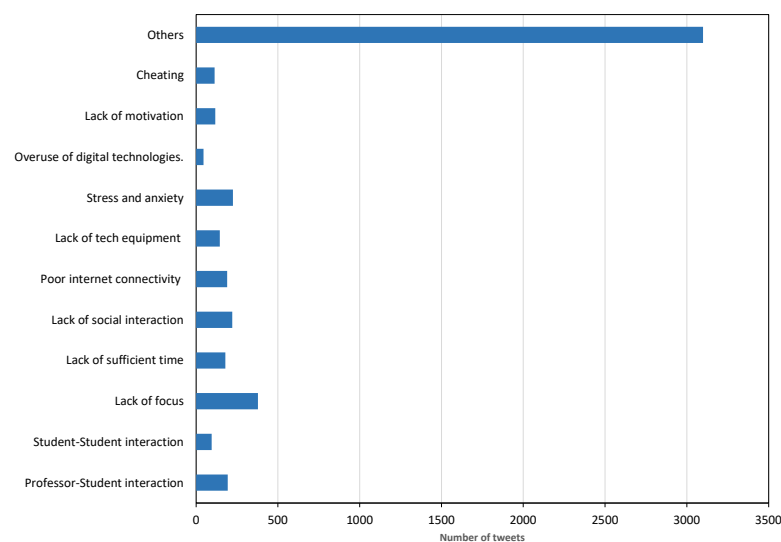
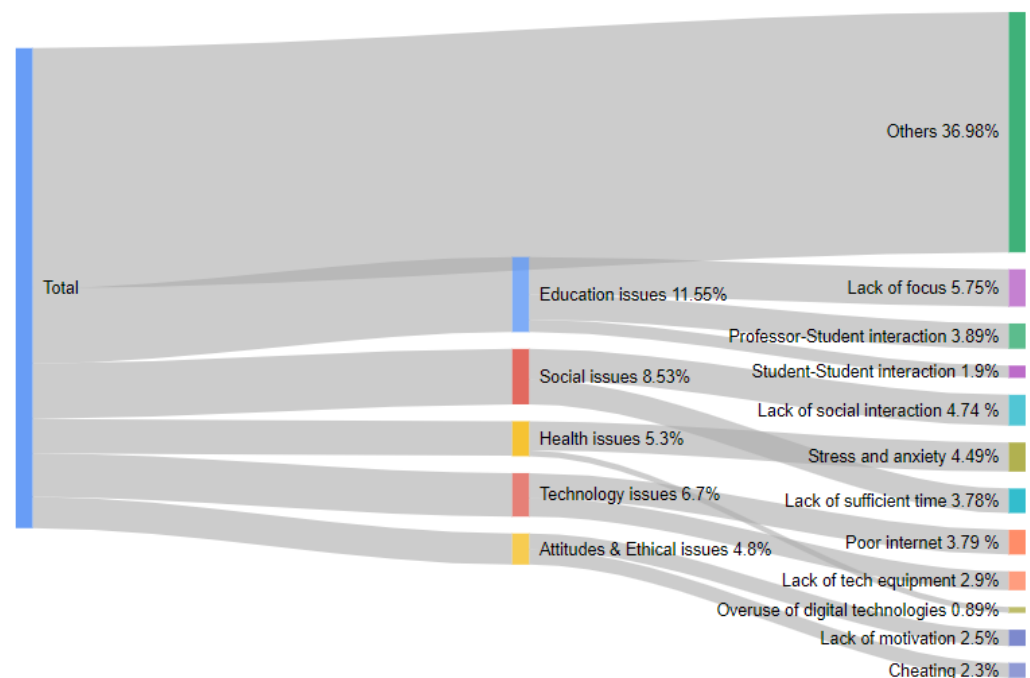


Figure 4. Number of tweets in each category.

Table 4. The counts and percentages of general and sub issues.

Issues	Tweets Count	Percentage
Education issues	23,976	11.56%
Professor–student interaction	8082	3.90%
Lack of focus	11,933	5.75%
Student–student interaction	3961	1.91%
Social issues	17,701	8.53%
Lack of time	7847	3.78%
Lack of social interaction	9854	4.75%
Technology issues	13,912	6.70%
Poor internet connectivity	7881	3.80%
Lack of tech equipment	6031	2.90%
Health issues	11,148	5.39%
Overuse of digital technologies	1864	0.90%
Stress and anxiety	9320	4.49%
Attitudes and ethical issues	9964	4.80%
Lack of motivation	5258	2.53%
Cheating	4706	2.27%
Others	130,751	63.01%

**Figure 5.** The distribution of general and sub issues.

4.2. Model Evaluation

The evaluation is performed on 70% for training and 30% for testing. To evaluate the performance of the model, we used the popular statistical measures: accuracy, precision, recall, and the harmonic average score (F1). F1 is a weighted average between precision and recall, where the best score is 1 and the worst is 0 [33]. Since our classification model is a multi-class classifier, we used the confusion matrix with multiple classes to measure the performance of the model [34]. An example of a confusion matrix focusing on the social issues (SI) class is shown in Table 5, where TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives, respectively.

Table 5. Confusion matrix for multi-class classification for social issues (SI) class vs. other classes: Education issues (EI), Social issues (SI), Technology issues (TI), Health issues (HI), and Attitudes and Ethical issues (HI).

Predicated Classes	Classes	EI	SI	TI	HI	AI	Others
Actual Classes	EI	TN	FP	TN	TN	TN	TN
	SI	FN	TP	FN	FN	FN	FN
	TI	TN	FP	TN	TN	TN	TN
	HI	TN	FP	TN	TN	TN	TN
	AI	TN	FP	TN	TN	TN	TN
	Others	TN	FP	TN	TN	TN	TN

The precision, recall, and F1 with respect to one specific category or sub-category are shown in Equations (3)–(5). Form the respective of all classes, micro-averaged and macro-averaged measures are used. The latter gives equal weight to all categories, whereas the former considers the different distributions of the categories. Equations (6)–(11) show precision, recall, and F1 for micro-average and macro-average [33], where μ and M indices represent micro- and macro-averaging.

$$Recall(R) = \frac{TP}{TP + FN} \quad (3)$$

$$Precision(P) = \frac{TP}{TP + FP} \quad (4)$$

$$F1 = \frac{2 * P * R}{P + R} \quad (5)$$

$$Recall_{\mu} = \frac{\sum_{i=1}^l TP_i}{\sum_{i=1}^l (TP_i + FN_i)} \quad (6)$$

$$Precision_{\mu} = \frac{\sum_{i=1}^l TP_i}{\sum_{i=1}^l (TP_i + FP_i)} \quad (7)$$

$$F1_{\mu} = \frac{2 * Precision_{\mu} * Recall_{\mu}}{Precision_{\mu} + Recall_{\mu}} \quad (8)$$

$$Recall_M = \frac{\sum_{i=1}^l \frac{TP_i}{TP_i + FN_i}}{l} \quad (9)$$

$$Precision_M = \frac{\sum_{i=1}^l \frac{TP_i}{TP_i + FP_i}}{l} \quad (10)$$

$$F1_M = \frac{2 * Precision_M * Recall_M}{Precision_M + Recall_M} \quad (11)$$

4.3. Classifier Result

From the category developing phase, we had a total of 5000 tweets annotated with five categories and 11 sub-categories. To train the classifiers, we used 70% for training (3500 tweets) and 30% for testing (1,500 tweets). We implemented a one-vs-all classifier using OneVsRestClassifier with LinerSVC where the kernel parameter is linear. We selected LinerSVC because it uses with multi-class one-vs-all mode. Hyper parameter C is tuned using an exhaustive grad search that considers all the parameters combinations. The linear kernel and C values (1, 10, 100, 1000) are estimated, and the best value for C was 10. Tables 6 and 7 demonstrate the result of the evaluation measures for the first and second-layer classifiers. The result of classifying the tweets into categories are summarized in Table 6, whereas Table 7 shows the result of classifying the tweets into more specific

issues. From Table 6, we note that the best result is achieved by educational issues and then social issues with F1 of 0.715 and 0.664, respectively. On other hand, the worst result is achieved by attitudes and ethical issues with an F1 of 0.616. We notice that the variation of the performance is because of the differences in the training instances used in each of the above categories, which are 456 and 233 for educational and social issues and 62 for attitudes and ethical issues. From Table 7, we find that best results are achieved by the topics related to social issues, which is closely followed by topics related to educational and technology issues. The worst results are achieved by topics related to attitudes and ethical issues. The detection of sub-category topics is satisfactory, with an F1 of 72%.

The results for the multi-layer multi-class classification model are summarized in Table 8. It is notable that the micro-averaged values register higher scores compared to macro-averaged due to the unbalanced class distributions. Thus, the more frequent the occurrence of the class, the more can be detected and classified accurately. Clearly, using the resulting category as a feature leads to improvement in the classification decision of the sub-category classifiers. The performance evaluation of the two-layer classifiers is shown in Figures 6 and 7; it can be seen that the precision, recall and F-measure of the sub-category classifier are around 72%, meanwhile, the category classifier is below 70%. If the classifier detects that a tweet belongs to “others”, then none of the lower-level classifiers would be applied.

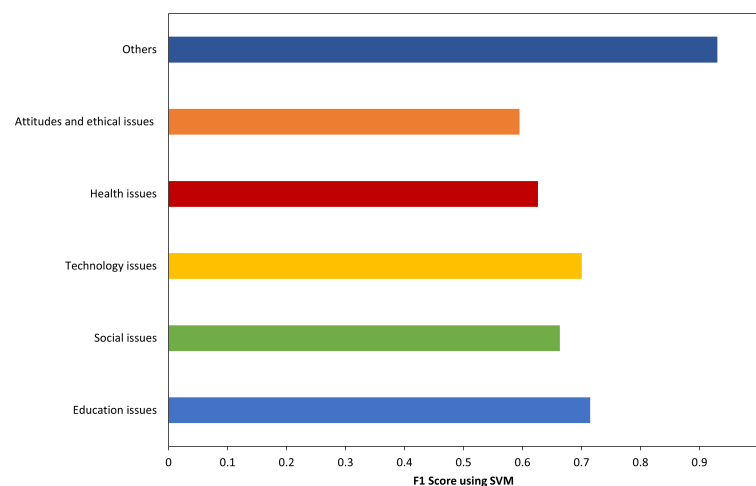


Figure 6. F1 score for each category.

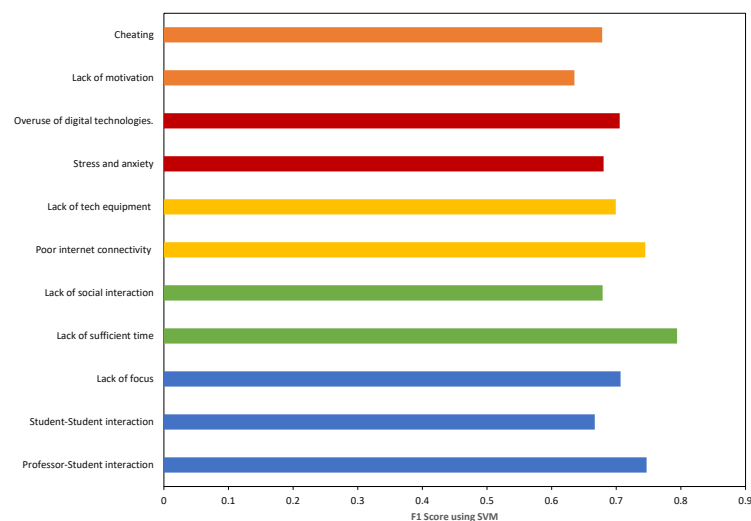


Figure 7. F1 score for each sub-category.

Table 6. Evaluation measures with SVM for each category.

Category	Precision	Recall	F1
Education issues	0.7262	0.7039	0.7149
Social issues	0.6776	0.6502	0.6636
Technology issues	0.6424	0.6644	0.6532
Health issues	0.6069	0.6476	0.6263
Attitudes and ethical issues	0.5357	0.7258	0.6164

Table 7. Evaluation measures with SVM for each sub-category.

Sub-Category	Precision	Recall	F1
Professor–student interaction	0.7284	0.7662	0.7468
Student–student interaction	0.6531	0.6809	0.6667
Lack of focus	0.7108	0.7023	0.7066
Lack of time	0.7941	0.7941	0.7941
Lack of social interaction	0.6223	0.7467	0.6788
Poor internet connectivity	0.7863	0.7077	0.7449
Lack of tech equipment	0.7049	0.6935	0.6992
Overuse of digital technologies	0.6346	0.7333	0.6804
Stress and anxiety	0.7176	0.6932	0.7052
Lack of motivation	0.662	0.61039	0.6351
Cheating	0.6941	0.6629	0.6782

Table 8. Macro-average and micro-average measures for multi-layer multi-class SVM model.

Issues	Macro-F1	Micro-F1
Category	0.6574	0.6788
Sub-category	0.7033	0.7209

5. Discussion and Limitation

Considering the results presented in the previous section, a detailed description of the discovered issues that represent students' experience in distance education will be presented in terms of analysis and comparison to the recent studies.

- **Education Issues:** Our analysis reveals that students find difficulties in adapting to a distance learning environment. The result shows that the highest proportion of tweets, around 11.55%, reflects problems such as the lack of interactions between students themselves and between students and teachers. Students who have been studying in physical classroom are finding difficulties focusing in online platforms. These issues might be raised because of the sudden change from face-to-face learning to a new model of learning. There are some students who lack communication skills and who are unprepared to interact in the class with their teachers and friends; they feel separated. This finding echoes a study on challenges in distance education [9], which indicates the lack of communication between teachers and students, and their finding shows that the teachers and students are not ready for distance education. In fact, with distance learning, the course plan and design should be aligned with the online learning style. Many studies [9,35,36] recommended to train teachers to be open to alternative ways of teaching.
- **Social Issues:** The analysis shows that 8.53% of tweets' content implies that social issues significantly influence student social well-being. Students trade off their daily life and social engagement to perform their studying tasks; they face difficulty adapting the new learning style, and that requires intensive work to manage their daily time. Social engagement provides positive impact and sense of predictability and stability [37]. Students might benefit from being involved in social activities because learning is a social process that is supported by many learning theories [38].
- **Technology Issues:** Technology is immensely useful in distance learning; several online platforms were available to support online education. Our analysis finds that technology problems such as internet shortage, technical problems, and shortage of digital equipment arise due to a deficiency of technology. Many students not equipped

with high-speed internet services, which plays an important role in how quickly they attend their classes. In addition, students may encounter technical issues as they lack skills related to using computer applications. All students could experience access issues in terms of problems with technology or lack of access [9,39]. In addition, there are some homes where there are many people with limited numbers of digital devices.

- **Health Issues:** Our analysis finds that students suffer from stress and anxiety problems. Moreover, online learning has increased the time students spend on digital devices, and that may impact on student's health. This radical change in learning style affects the lifestyle and can lead to loneliness, anxiety, and depression. Some studies have revealed that students are suffering from stress and anxiety during this pandemic period [40,41].
- **Attitude and Ethical Issues:** Another factor influencing students' success is their motivation and discipline. A significant problem in distance education is the ability of student to devote to their learning. A percentage of 4.8% of tweets' content is related to the motivation and discipline of students. Some students misunderstand the distance education; they deal with it as a vacation because they are at home through the learning process. This finding confirmed the results stated by [9] that students did not absorb the importance of the lessons because of their perceptions and lack of face-to-face communication. A lack of control leads students not to take their tasks seriously [42]. The issue of cheating is one of the most declared problems in online education during COVID-19. The most noted reason for cheating is the lack of proper monitoring systems. Teachers concerned about the risk of cheating feel that students might cheat to get better scores. Cheating in an online environment takes less effort compared to face-to-face [43]. According to [44], the perception of cheating of the majority of students is consistent with a sociologic perspective of deviance. Therefore, understanding students' perception of cheating in online learning is important for institutions and developers to raise the level of honesty in the academic environments.

By comparing our finding with recent studies, we found that education issues were dominant among other issues. In particular, a lack of focus was noted as the most significant educational challenge. This result ties well with the finding by Hietanen et al. [45] wherein the highest percentage of challenges for students was pedagogical challenges followed by technical challenges. In contrast, they found that the lack of personal contact was the highest pedagogical challenge which comes second in our finding. In addition, an agreement to our finding by [27] reveals students found interactions through the class challenging, and it is easy to lose focus.

The most common technological issues found by our study are lack of tech equipment and poor internet connection, which are consistent with what has been found in [9], in which the lack of infrastructure and unreliable internet connection are the largest challenges in technology. Furthermore, they showed that technological challenges are the main source of the problems that students meet in their distance education process. More interestingly, [45] found that a lack of technological devices is not an issue because the technical infrastructure was already in place. However, using the technological tools was the most technical challenge, which disagrees with our finding. Another study by Abuhammad [26] was broadly in line with our result that showed that the technical barriers was the most discussed theme in their data because of the insufficient maintenance and connectivity. In addition, we agree with Segbenya [28], who suggested that students must acquire online devices to ensure the success of their learning experience. Such lack of tech devices could negatively impact on distance education.

A finding by Khalil et al. [46] reveals that time organization was improved because online sessions save students' time. In addition, the students found enough time to spend with their family and friends. Our finding is almost the opposite of that: a lack of time and social interaction with family and friends has a negative impact on students' distance learning.

This study demonstrates that lack of motivation, stress, and anxiety are social issues that prevent students from pursuing their study effectively. Similar issues were obtained by Segbenya [28]; they found that challenges for students during online lectures were motivation and fear of collaboration. Another study by Kari et al. [27] supports our finding that students require more effort to stay motivated and have self-discipline.

Overall, most of the studies concluded that the most crucial issues among students were technological and education barriers. Additionally, a lack of face-to-face contact might cause more challenges such as lack of focus, lack of social interaction, and low motivation and well-being, which could also provoke a stressful situation.

The current findings offer a wide picture of the experience of students during their distance education. Students have increasingly become engaged in the improvement of their learning experience. Their feedback regarding the learning process has become an indicator of educational quality. The present materials in their social media content could provide insights in which educators can make improvements to the quality of the learning process. For example, conversations about lessons and material can give insight into how to interpret and develop the concepts of the lesson, which will impact on the quality of teaching and learning.

A large stream of Twitter posts that discuss sparsely events were annotated as 'others'. In this category, some tweets do not have obvious meaning; also, some do not reflect students' concerns. Since Twitter posts are considered as user-generated content, they follow the long-tail distribution [47]. Therefore, a small number of tweets that exhibit major students' problems have high occurrence, whereas the large numbers of tweets that exhibit insignificant problems have low occurrence. Investigation of this category can reveal insightful information.

Certain aspects of the work uncover some limitations in achieving the goal. Firstly, we found that most of tweets' content include issues and problems; on the other hand, several tweets contain positive aspects. However, the focus of this study is on the problems because they could be the most revealing and insightful for improvement of the education process. Secondly, great numbers of tweets were classified as 'Others' due to their lack of sufficient description. There are a variety of other issues hidden in this category, but this study did not conduct further analysis into the content of it. Lastly, the study is an exploratory study to discover the main issues related to distance education. The classification was built based on label independency. In fact, we did not consider the correlation between the issues under classification. However, to gain a full understanding would be optimal to address the correlations among these problems.

6. Conclusions

The informal conversations on social media are increasing steadily. Students' discussion on Twitter can provide valuable information to enlighten the learning process. The analysis of the content of such data is challenging and needs automatic data analysis techniques. This study designed a model for topics detection and classification in online distance education. We have proposed an automated methodology to analyze social media content with minimal human involvement. In summary, the unsupervised machine learning algorithm employed in this study detects themes that are important in identifying issues are relevant to distance education as clusters. Following this process, a sample of tweets are annotated to better evaluate the correlated content of the identified themes. Afterwards, SVM classifiers are trained to classify the tweet under consideration to one of the constructed categories. The results of this study show that the dominant issues that emerged were education and social issues. In addition, lack of interaction and poor internet connections affect the successful learning. The result could help the future studies to investigate in depth the students' perspectives and concerns on different topics associated with their educational process. The perceived issues provide practical reference for improving the quality of education and gain further understanding of students' educational experience. Future work is a follow-up research focusing on exploring the

correlation between the topics to improve the accuracy of the model. Furthermore, we could investigate the “Others” content to detect the hidden low-frequency themes to reveal more insightful information.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to Twitter’s Terms and Conditions.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BTM	Biterm topic model
SVM	Support Vector Machine
PLSI	Probabilistic semantic based indexing model
LDA	Latent dirichlet allocation
KSA	Kingdom of Saudi Arabia

References

- Huang, R.; Tlili, A.; Chang, T.; Zhang, X.; Nascimbeni, F.; Burgos, D. Disrupted classes, undisrupted learning during COVID-19 outbreak in China: application of open educational practices and resources. *Smart Learn. Environ.* **2020**, *7*, 19. <https://doi.org/10.1186/s40561-020-00125-8>.
- Jena, P.K. Impact of Pandemic COVID-19 on Education in India. *Int. J. Curr. Res. (IJCR)* **2020**, *12*, 12582–12586.
- Oyediran, W.; Ayodeji Motunrayo, O.; Owoyemi, M.; Adejobi, A.; Fasasi, R. Prospects and limitations of e-learning application in private tertiary institutions amidst COVID-19 lockdown in Nigeria. *Heliyon* **2020**, *6*, e05457. <https://doi.org/10.1016/j.heliyon.2020.e05457>.
- Mahyoob, M. Challenges of e-Learning during the COVID-19 Pandemic Experienced by EFL Learners. *Arab World Engl. J.* **2020**, *11*, 351–362. <https://doi.org/10.24093/awej/vol11no4.23>.
- Sathishkumar, V.; Radha, R.; Saravanakumar, A.; Mahalakshmi, K. E-Learning during Lockdown of Covid-19 Pandemic: A Global Perspective. *Int. J. Control Autom.* **2020**, *13*, 1088–1099.
- Hoq, M. E-Learning During the Period of Pandemic (COVID-19) in the Kingdom of Saudi Arabia: An Empirical Study. *Am. J. Educ. Res.* **2020**, *8*, 457–464.
- Adedoyin, O.B.; Soykan, E. Covid-19 pandemic and online learning: the challenges and opportunities. *Interact. Learn. Environ.* **2020**, 1–13. <https://doi.org/10.1080/10494820.2020.1813180>.
- Armstrong-Mensah, E.; Ramsey-White, K.; Yankey, B.; Self-Brown, S. COVID-19 and Distance Learning: Effects on Georgia State University School of Public Health Students. *Front. Public Health* **2020**, *8*, 547. <https://doi.org/10.3389/fpubh.2020.576227>.
- Sari, T.; Nayir, F. Challenges in Distance Education During the (Covid-19) Pandemic Period. *Qual. Res. Educ.* **2020**, *9*, 328–360. <https://doi.org/10.17583/qre.2020>.
- Altwaijry, N.; Ibrahim, A.; Binsuaidan, R.; Alnajjar, L.; Alsouk, B.; Almutairi, R. Distance Education During COVID-19 Pandemic: A College of Pharmacy Experience. *Risk Manag. Healthc. Policy* **2021**, *14*, 2099–2110.
- Suda, K.; Sterling, J.; Guirguis, A.; Mathur, S. Student perception and academic performance after implementation of a blended learning approach to a drug information and literature evaluation course. *Curr. Pharm. Teach. Learn.* **2014**, *6*, 367–372. <https://doi.org/10.1016/j.cptl.2014.02.017>.
- Elfirdoussi, S.; Lachgar, M.; Kabaili, H.; Rochdi, A.; Goujdami, D.; El Firdoussi, L. Assessing Distance Learning in Higher Education during the COVID-19 Pandemic. *Educ. Res. Int.* **2020**, *2020*, 8890633. <https://doi.org/10.1155/2020/8890633>.
- Siemens, G.; Long, P. Penetrating the Fog: Analytics in Learning and Education. *Educ. Rev.* **2011**, *5*, 30–32. <https://doi.org/10.17471/2499-4324/195>.
- Hofmann, T. Probabilistic Latent Semantic Indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA USA, 15–19 August 1999; SIGIR ’99; Association for Computing Machinery: New York, NY, USA, 1999; pp. 50–57. <https://doi.org/10.1145/312624.312649>.
- Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. <https://doi.org/10.1162/jmlr.2003.3.4-5.993>.
- Nagori, R.; Gnanasekaran, A. LDA based integrated document recommendation model for e-learning systems. In Proceedings of the 2011 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Udaipur, India, 22–24 April 2011; pp. 230–233. <https://doi.org/10.1109/ETNCC.2011.6255892>.

17. Zhong, J.; Zhang, S.F.; Guo, W.L.; Li, X. TFLA: A Quality Analysis Framework for User Generated Contents. *Tien Tzu Hsueh Pao/Acta Electron. Sin.* **2018**, *46*, 2201–2206. <https://doi.org/10.3969/j.issn.0372-2112.2018.09.022>.
18. Yan, X.; Guo, J.; Lan, Y.; Cheng, X. A biterm topic model for short texts. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, May 13–17 2013; ACM: New York, NY, USA, 2013; pp. 1445–1456. <https://doi.org/10.1145/2488388.2488514>.
19. Romero, C.; Ventura, S. Educational Data Mining: A Review of the State of the Art. *Syst. Man Cybern. Part C* **2010**, *40*, 601–618. <https://doi.org/10.1109/TSMCC.2010.2053532>.
20. Kotsiantis, S.; Pintelas, P. Predicting Students Marks in Hellenic Open University. In Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05), Kaohsiung, Taiwan, 5–8 July 2005; Volume 2005, pp. 664–668. <https://doi.org/10.1109/ICALT.2005.223>.
21. van Rooyen, A. Distance Education Accounting Students' Perceptions of Social Media Integration. *Procedia Soc. Behav. Sci.* **2015**, *176*, 444–450. <https://doi.org/10.1016/j.sbspro.2015.01.495>.
22. Samigulina, G.; Samigulina, Z. Intelligent System of Distance Education of Engineers, Based on Modern Innovative Technologies. *Procedia Soc. Behav. Sci.* **2016**, *228*, 229–236. <https://doi.org/10.1016/j.sbspro.2016.07.034>.
23. Kumar, A.; Murugappan, A. A multi-aspect based Opinion Mining System for Open and Distance Education using Online Reviews. *Online J. Distance Educ. -Learn.* **2017**, *5*, 1–11.
24. Aljabri, M.; Chrouf, S.M.; Alzahrani, N.A.; Alghamdi, L.; Alfehaid, R.; Alqarawi, R.; Alhuthayfi, J.; Alduhailan, N. Sentiment Analysis of Arabic Tweets Regarding Distance Learning in Saudi Arabia during the COVID-19 Pandemic. *Sensors* **2021**, *21*, 5431.
25. Alqurashi, T. Stance Analysis of Distance Education in the Kingdom of Saudi Arabia during the COVID-19 Pandemic Using Arabic Twitter Data. *Sensors* **2022**, *22*, 1006.
26. Abuhammad, S. Barriers to distance learning during the COVID-19 outbreak: A qualitative review from parents' perspective. *Heliyon* **2020**, *6*, e05482.
27. Almendingen, K.; Morseth, M.S.; Gjølstad, E.; Brevik, A.; Tørris, C. Student's experiences with online teaching following COVID-19 lockdown: A mixed methods explorative study. *Scand. J. Educ. Res.* **2021**, *16*, e0250378.
28. Segbenya, M.; Bervell, B.; Minadz, V.M.; Somuah, B.A. Modelling the perspectives of distance education students towards online learning during COVID-19 pandemic. *Smart Learn. Environ.* **2022**, *9*, 13.
29. Terpilowski, M. Bitermplus 2021. Available online: <https://bitermplus.readthedocs.io/en/latest/tutorial.html> (accessed on 20 April 2022).
30. Rosner, F.; Hinneburg, A.; Röder, M.; Nettling, M.; Both, A. Evaluating Topic Coherence Measures. 2013. Available online: https://www.researchgate.net/publication/261101181_Evaluating_topic_coherence_measures (accessed on 20 April 2022).
31. Bouazizi, M.; Ohtsuki, T. Multi-class sentiment analysis on twitter: Classification performance and challenges. *Big Data Min. Anal.* **2019**, *2*, 181–194. <https://doi.org/10.26599/BDMA.2019.9020002>.
32. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al. API design for machine learning software: Experiences from the scikit-learn project. *arXiv* **2013**, arXiv:1309.0238.
33. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>.
34. Grandini, M.; Bagli, E.; Visani, G. Metrics for Multi-Class Classification: An Overview. *arXiv* **2020**, arXiv:2008.05756.
35. Arinto, P. Issues and Challenges in Open and Distance e-Learning: Perspectives from the Philippines. *Int. Rev. Res. Open Distrib. Learn.* **2016**, *17*, 162–180. <https://doi.org/10.19173/irrodl.v17i2.1913>.
36. Pant, A. Distance Learning: History, Problems and Solutions. *Adv. Comput. Sci. Inf. Technol. (ACSIT)* **2014**, *1*, 65–70.
37. Cohen, S.; Wills, T. Stress, Social Support, and the Buffering Hypothesis. *Psychol. Bull.* **1985**, *98*, 310–57. <https://doi.org/10.1037/0033-2909.98.2.310>.
38. Pritchard, A.; Woollard, J. *Psychology for the Classroom: Constructivism and Social Learning*; Routledge: London, UK, 2010.
39. Fey, S.; Emery, M.; Flora, C. Student issues in distance education programs: do inter-institutional programs offer students more confusion or more opportunities? *J. Asynchronous Learn. Netw.* **2019**, *12*, 71–83.
40. Chakraborty, P.; Mittal, P.; Gupta, M.; Yadav, S.; Arora, A. Opinion of students on online education during the COVID-19 pandemic. *Hum. Behav. Emerg. Technol.* **2020**, *3*, 357–365. <https://doi.org/10.1002/hbe2.240>.
41. Islam, M.; Dey Barna, S.; Raihan, H.; Khan, M.; Hossain, M. Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: A web-based cross-sectional survey. *PLoS ONE* **2020**, *15*, e0238162. <https://doi.org/10.1371/journal.pone.0238162>.
42. Khoury, A.; Nasir Eddeen, L.; Saadah, D.; Harfoushi, O. E-Learning: Justifications and Obstacles. *ijET* **2011**, *6*, 53–56. <https://doi.org/10.3991/ijet.v6i3.1610>.
43. Nagi, K. Solving Ethical Issues in eLearning. *Int. J. Comput. Internet Manag.* **2006**, *14*, 7.
44. Raines, D.A.; Ricci, P.; Brown, S.L.; Eggenberger, T.; Hindle, T.; Schiff, M. Cheating in online courses: The student definition. *J. Eff. Teach.* **2011**, *11*, 80–89.
45. Hietanen, M.; Svedholm-Häkkinen, A.M. Transition to Distance Education in 2020—Challenges among University Faculty in Sweden. *Scand. J. Educ. Res.* **2022**, 1–14. <https://doi.org/10.1080/00313831.2021.2021444>.

-
46. Khalil, R.; Mansour, A.E.; A.Fadda, W.; Almisnid, K.; Aldamegh, M.; Alkhalifah, A.A.N.A.; Al-Wutayd, O. The sudden transition to synchronized online learning during the COVID-19 pandemic in Saudi Arabia: a qualitative study exploring medical students' perspectives. *BMC Med. Educ.* **2020**, *20*, 285.
 47. Ochoa, X.; Duval, E. Quantitative Analysis of User-Generated Content on the Web. 2008. Available online: https://www.researchgate.net/publication/228981330_Quantitative_analysis_of_user-generated_content_on_the_Web (accessed on 20 April 2022).