

Article

Analysis of Learner's Sentiments to Evaluate Sustainability of Online Education System during COVID-19 Pandemic

Muhammad Imran ^{1,*}, Saman Hina ^{1,*}  and Mirza Mahmood Baig ²

¹ Department of Computer Science & IT, NED University of Engineering and Technology, Karachi 75270, Pakistan; hmimran@neduet.edu.pk

² Department of Mathematics, NED University of Engineering and Technology, Karachi 75270, Pakistan; cdm@neduet.edu.pk

* Correspondence: samhaque@neduet.edu.pk

Abstract: Education is an important domain that may be improved by analyzing the sentiments of learners and educators. Evaluating the sustainability of the education system is critical for the continuous improvement and satisfaction of the learner's community. This research work focused on the evaluation of the effectiveness of the online education system that has been adopted during the COVID-19 pandemic. For this purpose, sentiments/reviews of learners were collected from the Twitter website regarding the education domain during COVID-19. To automate the process of evaluation, a hybrid approach was applied that used a knowledgebase of opinion words along with machine learning and boosting algorithms with n-grams (unigram, bigram, trigram and combination of all these n-grams). This automated approach helped to evaluate the transition of the education system in different circumstances. An ensemble classifier was created in combination with a customized knowledgebase using classifiers that individually performed best with each of the n-grams. Due to the imbalanced nature of the data (tweets), these operations were performed by applying the synthetic minority oversampling technique (SMOTE). The obtained results show that the use of a customized knowledgebase not only improved the performance of the individual classifiers but also produced quality results with the ensemble model. As per the observed results, the online education system was not found sustainable as the majority of the learners were badly affected due to some important aspects (health issues, lack of training and resources).

Keywords: sustainability; online education; ensemble; machine learning; sentiment analysis; SMOTE



Citation: Imran, M.; Hina, S.; Baig, M.M. Analysis of Learner's Sentiments to Evaluate Sustainability of Online Education System during COVID-19 Pandemic. *Sustainability* **2022**, *14*, 4529. <https://doi.org/10.3390/su14084529>

Academic Editor: Rosabel Roig-Vila

Received: 28 January 2022

Accepted: 7 April 2022

Published: 11 April 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

Due to the COVID-19 pandemic, the shift from on-site education systems to online education systems has affected the quality of education. To evaluate the effectiveness of this changeover and to analyze the factors that may affect the sustainability of this transition, feedback from the learners is crucial. Feedback that is available on the internet in the form of reviews, tweets and blogs contains some hidden and unstructured information. This unstructured information can be processed using natural language processing techniques to reveal important aspects of a matter under consideration. Sentiment analysis (SA) and aspect-based sentiment analysis (ABSA) are among the more common methods to uncover the hidden facts behind the opinions of people about an entity. Nowadays, a large number of organizations collect customers' feedback data from social media websites to know people's responses, views, opinions or emotions about their products or services. Analysis of these reviews using the sentiment analysis technique helps in strategic decision making to ensure sustainable growth in the business. This approach has become even more useful due to the drastic changes in the global business environment brought about by the COVID-19 pandemic. In addition, an automated framework to perform sentiment analysis may help organizations to analyze customer feedback in an efficient manner. This

is because the associated data (tweets, blogs, reviews, etc.) are usually enormous in size, and therefore, it is a hectic and time-consuming task to manually analyze them.

There are several classical machine learning and deep learning methods available that are very useful for the analysis of sentiment polarity toward any product, service, global event, etc. [1,2]. The present study focused on developing a hybrid framework to detect and classify the polarity toward online education during the COVID-19 pandemic. This hybrid approach also utilized SMOTE for tackling the data imbalance issue. In addition, we also developed an ensemble classifier in combination with a customized knowledgebase to improve the performance of the automated framework. It was envisaged that the outcome of this study would highlight the crucial factors to be considered to make online teaching and learning effective and sustainable in the long term.

2. Literature Review

During the COVID-19 pandemic, the global education system faced a drastic transition in terms of teaching and learning from the on-site to online mode. A related study stated that this change had been risky to bridge the gap between the teaching and learning communities [3]. This risk is not only related to the health and safety of an individual who is restricted to working/learning from home but to technical issues as well. This has led to the long-term sustainability issue in the case of a sudden change in the existing system. Moreover, it was found challenging for learners to adapt to this new learning mechanism with limited resources. Adaptation to change in the education sector was difficult for any group of learners, and it was equally critical to evaluate the competitiveness of newly adapted methods (i.e., online education). In addition to this change, to make online education sustainable and adaptable, the evaluation of this new mode of teaching and learning is also crucial. This evaluation can highlight important aspects (lack of resources, different learning styles, etc.) that may be used to improve the quality of online education. In this respect, a group of researchers has proposed a hybrid framework incorporating convolutional neural networks (CNNs) and a gated recurrent unit (GRU)-based recurrent neural network for the detection of the learning style of learners [4]. This did not target sentiments of the learner's community. In this pandemic condition, most of the institutions were closed, and the teaching shifted toward virtual learning. A study was conducted to uncover a digital divide in this situation and how this change affected the students' academic performance [5]. Another study was carried out that analyzed the opinions of teachers, students and their parents to identify the psychological effects of social isolation on students and the continuation of virtual education. For this study, a questionnaire was designed for middle school, high school and university students. The results of this study show that online study is not as effective as traditional education but is effective and sustainable in a situation like the current one [6].

Another source to evaluate important aspects in the education sector is through collection of feedback/reviews from students/learners. For instance, in the education sector, most institutions collect feedback from the students at the end of the session/course/workshop to measure the quality of teaching, syllabus and the resources available for delivering the lecture. Working on this feedback/reviews helps to improve the quality of education in terms of learning and resources. Standard systems of feedback analysis were serving well for the past many years, but the uncertain situation of the COVID-19 pandemic affected the education system badly. Unfortunately, the COVID-19 pandemic adversely affected not only public health but also the economic, cultural, religious and learning structures of most countries around the globe. At this time, as the world has been facing the COVID-19 pandemic for several months now, many institutions in the world have shifted or are planning to shift toward online learning, whereas other institutions, particularly in developing countries, are unable to shift toward online learning or have closed due to unavailability of online learning resources (proper electric supply, internet facilities, electronic gadgets). According to one research study, approximately 465 M children around the globe are unable

to continue their education due to improper facilities and lack of resources [7]. Due to this, parents and educators are concerned about the future of students.

After COVID-19, teaching all over the world moved online [8]. Despite the negative impact of the pandemic, using technology is cost-effective, advanced and practical. Research found that unavailability of learning resources, family poverty, inadequate IT infrastructure and health concerns (due to continuous usage of electronic gadgets) are the main areas that required attention for effective online learning [9].

Due to the frequent growth of data on the internet, the importance of sentiment analysis is increasing day by day as it can uncover hidden facts. Majority of the work related to sentiment analysis (analysis of reviews/feedback of customers/users) has been reported against products (mobile, laptop, etc.) or services (e.g., restaurants). There are different ways to analyze people's feedback or views regarding an entity. Presently, it is the most impactful domain of interest to computer science researchers. For identification of aspects from a learner's sentiments/feedback, Wang et al. [10] proposed gradual machine learning (GML) that performs better than its unsupervised substitute without requiring human effort for labeling the data. Pinlong Zaho et al. [11] proposed a different approach for the analysis of sentiments/reviews using online tweets. Another approach transformed emoticons and emojis into plain text and then applied bidirectional encoder representations from transformers (BERTs), trained on plain text, to classify tweets [12]. A supervised model for aspect-based sentiment analysis (ABSA) uses two long short-term memory (LSTM) layers. The first layer predicts the aspects and the second layer classifies the polarity of the predicted aspects. This analysis attained a good accuracy result using five years of student feedback. This study was the first attempt to perform an aspect-based sentiment analysis of students' feedback using a neural network model [13]. The main hurdle of online learning is the availability of required resources and facilities. A statistical study found that the education system of developing countries was seriously damaged by the COVID-19 pandemic. The reason is that these countries were not equipped for online education, and procurement of this new technology has many challenges. The reported study mainly emphasized the difficulties of online education and suggested solutions to survive any further crises in the future [7]. Many researchers used semi-supervised models for ABSA tasks.

Because there are few studies reported related to the education domain, we also considered methods to analyze sentiments for other domains (products, services). For instance, Ning Li et al. presented a semi-supervised multi-task learning (SEML) framework with extensive experiments on ABSA using four review datasets of SemEval workshops [14]. This study performed an ABSA of student opinion surveys in the Serbian language [15]. Nandal, N. et al. proposed a novel approach for detecting aspect-level sentiments using Amazon customer reviews. The reviews were first pre-processed (stemming, tokenization, casing, stop-word removal) to extract useful results [16]. This research proposed lexicon generation methods (statistical and genetic algorithm) for aspect-based problems. Experiments showed that this method outperforms baseline methods in aspect-based polarity classification of Bing Liu's customer review datasets with an improved F-measure, precision and recall [17]. Liu, N et al. presented a gated alternate neural network (GANN), which is language-independent and achieved state-of-the-art results [18]. Another study constructed a domain-specific sentiment lexicon comprising two steps: The first selected product features and sentiment words to find relationships between them using an improved algorithm. In the second step, sentiment words related to the domain (mobile shopping) were clustered and categorized into sentiment dimensions. This new domain-specific lexicon was evaluated using other product reviews in the Chinese and English languages by machine learning and deep learning models, and the results are outstanding and show the effectiveness of this approach [19]. Another study used a weakly supervised method based on a convolutional neural network (CNN) to identify keywords distinguishing positive and negative sentences in movie review datasets in which every word was represented as a continuously valued vector and each sentence was represented as a matrix.

The scores for classification of sentence polarity and identification of words were high [20]. In another study, a combined targeted aspect detection and aspect polarity classification was performed, and its performance was evaluated on two benchmark datasets. The experiment showed that these joint operations performed well on their dataset [21]. Moreover, a novel multi-layer architecture to represent customer reviews was proposed that showed that the sentiment for a product is composed of its aspects, expressed in sentences.

To classify the sentiments in product reviews, a hybrid of random forest (RF) and SVM was proposed [22]. Compared with other methods that were employed individually, it was found that the hybrid approach can increase the proficiency of sentiment analysis. Ditiman Hazarika et al. [23] used TextBlob (Python library for lexicon-based sentiment analysis) for the classification of tweets into three polarities: positive, negative or neutral. In terms of feature extraction that adds value to classifiers, several research studies have been reported. For instance, word embeddings also contribute to contextual understanding of sentiments. Muhammad Imran et al. [24] published the first largest word2vec embedding trained on 52 million tweets related to crises. They conducted a research study to validate the usefulness of human-annotated data. The data comprised 19 different crises that took place from 2013 to 2015. Similarly, term frequency–inverse document frequency (TF-IDF) is a method that is widely used in research to extract text features particularly for sentiment analysis tasks. Ravinder Ahuja et al. [25] performed sentiment classification of an SS-Tweet dataset using n-grams and TF-IDF on different classifiers. Results reveal that the logistic regression classifier performed the best compared with other classifiers, and it was concluded that machine learning algorithms produced the best results with TF-IDF features. Researchers also proposed two models to classify product reviews [26]. Model 1 implemented POS tagging, and model 2 used TF-IDF. Model 1 extracted aspect terms, whereas model 2 extracted features. They used Naïve Bayes, SVM, decision tree and KNN for sentiment classification. Experimental results indicate that SVM and Naïve Bayes performed similarly. Another research study [27] applied improved TF-IDF that replaced the traditional method for calculating IDF. This method involved the place coefficient and POS for calculating the weight. The results show that this attained good results. TF-IDF was also employed in another study that implemented latent semantic analysis (LSA) [28].

WordNet is another rich source that may help in the customization of available lexicons to capture more sentiments in the dataset. Each WordNet synset [29] is related to some scores that describe how positive, negative and objective terms are allied in the synset. On the other hand, Marco Guerini et al. [30] proposed an approach that resolved the posterior-to-prior polarity problem. They used two different versions of SentiWordNet, and their proposed approach performed well on different datasets in regression and classification tasks. Slangs are shorthand words that people use on the internet, especially social media blogs and product reviews. These slangs have to be replaced with proper words for effective sentiment classification. SentiWordNet lexicons can be used to identify slangs. In [31,32], a framework was presented that detected and scored slang words for opinion mining. The proposed approach achieved adequate results on micro blog datasets.

It is evident from the literature that lexicon-based methods comprising sentiment lexicons are very useful for performing opinion mining tasks since they do not require any type of training data to evaluate the results. SentiWordNet (SWN) is an extensively used lexicon for this task. SentiWordNet depends upon glosses [33], and research that combines SentiWordNet with other approaches may yield better results than individual classifiers.

In Ref. [34], it was reported that 90% of the terms in SentiWordNet are categorized as objective, and these terms are unable to provide useful information for classification. This incompleteness leads to the development of machine learning or a hybrid of machine learning and lexicon-based classifiers. Moreover, lexicon-based methods have some limitations:

1. They are challenging when extracting sentiments from a dynamic environment due to a limited number of words in lexicons.
2. Sentiment lexicons dispense sentiments based on their scores, deprived of their context in the sentence as words' sentiments differ from context to context.

In the education domain, a useful study was conducted by Kastrati, Z. et al. to analyze students' opinions on MOOCs. The goal of this framework was to automate the sentiment analysis against a given aspect associated with the MOOCs. This proposed model reduces the requirement of manually annotated data and achieves good results for aspect category and sentiment classification. The dataset for this research was students' reviews gathered from Coursera [35]. Word embedding can also help to improve performance. One study analyzed sentiments that used word2vec, CNNs and hyper parameters to tune CNNs with the genetic algorithm (GA). Results show that the proposed method performs better than other techniques with 95.5% accuracy [36]. Da'u, A., et al. proposed a system, aspect-based opinion mining (ABOM), that uses the deep learning technique to improve accuracy. The proposed model used a multichannel deep convolutional neural network (MCNN) and tensor factorization (TF). Compared with the baseline model, the model proposed by this study achieved significant results [37]. Various deep learning techniques have been used for aspect-based sentiment analysis [38]. A new framework for ABSA that used the whole context, lexicon embedding and the attention module was proposed for intricate syntactic dependency between words. The experiment was performed on four benchmark datasets and achieved effective results [39]. Xing, Y. et al. introduced a CNN model for aspect-level sentiment classification using a Twitter dataset, and it was observed that the CNN improves performance [40]. A document, tweet or review may contain implicit or explicit aspects in it. A research study [41] was conducted to extract implicit aspects from a document through co-occurrence and ranking.

The model proposed by these researchers uses aspects and sentiment words. First, they developed a deep learning model to classify sentiments using word embedding and a linear machine learning algorithm as a baseline to compare results. Second, two ensemble techniques were proposed that aggregate their baseline classifier with other surface classifiers widely used in sentiment analysis. In addition to this, two models were implemented for combining surface and deep features to merge information from several sources. Fourth, they introduced a taxonomy to classify various models and found in the literature, as well as the ones we proposed. Fifth, they performed experiments and compared performance with deep learning models. For this, they used seven public datasets that were extracted from the micro blogging and movie reviews domain. Obtained results show improved performance on the F1-Score [42]. Another aspect-based sentiment analysis was performed on a restaurant review dataset that used a convolutional neural network with a labeled dataset and word2vec as input word embeddings to identify the aspect category and sentiment polarity of the extracted aspect category. The results of this research study show that fine-tuned word embeddings performed well [43]. A hybrid fine-grained aspect-based sentiment analysis was performed by Zainuddin, N. et al. on tweets. This approach produced meaningful results with improved accuracy performance by 76.55, 71.62 and 74.24% [44].

Applying this observation, a multi-layer was designed to predict the ratings of a product. This model generates aspect ratings and weights and provides better results than the methods previously used [45]. A multi-layer CCN was studied that used word embeddings (word2vec and GloVe) and one-hot character vectors. It was found that the proposed model achieved state-of-the-art results in the detection of aspect category and aspect sentiment classification [46]. Shervin Minaee et al. [47] conducted a detailed study on the performance of deep learning models on a benchmark dataset to classify text. Other researchers proposed a model that can be used for labeling without any human effort. It was shown that the performance of the proposed model was better than using an unsupervised model. A knowledgebase model was proposed based on a knowledge graph to analyze social network sentiments. The efficiency of the model was measured with various metrics [48]. Nikola Nikolic et al. [15] introduced a system to analyze student feedback in the Serbian language using machine learning models, dictionaries and rules.

In contrast to the available research studies, it has been observed that facts derived from user reviews play an important role in the evaluation of any product, service or

system. To the best of our knowledge, none of the reported studies reported a hybrid approach comprising an ensemble learning classifier in combination with a customized knowledgebase. Moreover, deep learning methods can perform better on large datasets and were not employed in this study due to the limited size of the dataset. The scalability of this framework will evaluate sentiments expressed by a learner's community that will be helpful in devising a strategy to adopt sustainable processes based on the circumstances (such as COVID-19). As per the understanding of the current studies, these facts may also contribute to the evaluation of an online education system and identification of important aspects that will be helpful in developing policies to make the education system sustainable.

3. Data and Methodology

This section explains the collection of relevant data from social media platform regarding online education during COVID-19 pandemic. In addition to this, details of noise in the dataset and its pre-processing are also discussed which is an essential phase for the application of hybrid approach. Then, the proposed hybrid approach and its evaluation against other methods are explained in detail.

3.1. Data Collection and Preparation

The availability of dataset for this research was lacking due to ethical requirements. Some organizations provided benchmark datasets over the internet for study purpose, but they were related to specific products and services. The dataset selected for this research work comprised English language tweets extracted from Twitter in a comma-separated file (.csv) using some specific keywords such as: education during COVID-19, effect of COVID-19 on education, etc. The goal of this research work was to analyze the public views on education during COVID-19. Thus, the tweets relevant to the education domain and its problems during COVID-19 were extracted for analysis. Total of 17,003 tweets were collected, from which all re-tweets, tweets containing URLs, incomplete tweets and all irrelevant tweets were removed. The dataset was then labeled manually using annotation guidelines that were developed for this research. Annotation of data is mandatory in case of applying supervised machine learning classifiers. In addition to this, labeled/annotated information also serves in the evaluation of applied methods. Supervised machine learning needs annotated data for training and evaluation. Noisy and erratic annotation may result in lower performance of the algorithm. The dataset for this study was annotated manually [24]. Even though it is time-consuming and requires human efforts, it is found that the best performance is achieved with manually annotated datasets [49]. Annotation process followed by the preparation of gold standard dataset is explained in Section 3.1.1.

3.1.1. Preparation of Gold Standard Dataset for Training and Evaluation

Usually, data are labeled/annotated by domain experts, using annotation guidelines. In this study, the dataset was annotated independently by three undergraduate students of the authors' university who have sufficient knowledge and proficiency in English language. Annotators were asked to annotate the dataset with two labels (sentiment polarity) which are given as "Sentiment Polarity: $P = \{\text{Positive and Negative}\}$ ". Annotation guidelines that were developed for this study are as follows:

1. Identify the language used in tweets. Annotate the tweets that are in "ENGLISH" language and ignore all other tweets that are not in English language.
2. Consider both English words present in tweet to make decision about sentiment expressed in the tweets. Each tweet has to be assigned a polarity, from a set $P = \{\text{Positive, Negative}\}$.
3. If the expression about a target in tweet is in favor, support, positive attitude or forgiveness, then it should be considered as Positive sentiment.
4. If the expression about a target in tweet shows condemnation, negative attitude, judgment, questioning and negative emotion, then consider its sentiment as Negative.
5. If the tweet shows both negative and positive attitudes, then leave it.

Using these annotation guidelines annotators annotated two sentiment labels (positive and negative) in the dataset. In case the tweet expressed multiple sentiments, we labeled the tweet that expressed the comprehensive sentiment polarity of the tweet. To scrutinize the consistency and reliability of annotation, we used statistical analysis and annotation agreement methods.

Annotation guidelines were first tested and verified over small sample that also served as training experience for the annotators to improve inter-annotator agreement [50]. Trained annotators obtained more agreement than untrained annotators. They were also advised to specify sentiment that is more suitable to the tweets. Approximately 94% of labels (as tabulated in Table 1) were according to the predefined guidelines. Only few were conflicted, which were resolved after talk and mutual settlement with the annotators and the authors.

Inter-Annotator Agreement (IAA) Metrics

There are several IAA methods that are used to assess the agreement among independent annotators. The percent agreement is the simplest measure to find the agreement between coders. The drawback of percent agreement is that it does not look into the agreement by chance. The pair-wise percentage agreement of annotators for this study is tabulated in Table 1.

Table 1. Percentage agreement.

Annotator Pair	Positive	Negative
1–2	0.95	0.92
1–3	0.95	0.93
2–3	0.94	0.94

Cohen’s kappa is widely used measure [51] due to its ease and robustness. It is a metric that is used to measure the agreement between two annotators using Equation (1). The pair-wise agreement of annotators using Cohen’s kappa is shown in Table 2.

$$K = \frac{A_o - A_e}{1 - A_e} \quad (1)$$

where:

A_o is observed agreement and

A_e is the measure of the agreement by chance of each pair of annotators to calculate the agreement.

Table 2. Cohen’s k score.

Annotator Pair	A_o	A_e	K
1–2	0.945	0.550	0.878
1–3	0.948	0.552	0.883
2–3	0.946	0.557	0.880

After agreement measure, tweets that were tagged differently or conflicted were discussed with the annotators [52]. It is evident from the literature that the kappa statistic shows that K score of 0.81 to 1.00 is “almost perfect” for classification [53], but in order to build a gold standard dataset, the ambiguity was resolved by obtaining the views of annotators on the disagreements. After resolving disagreed annotations, the dataset was finalized to use with machine learning classifiers for sentiment classification. The distribution of polarities was identified as shown in Table 3.

Table 3. Count of polarities in the dataset.

Number of Tweets		
Negative	Positive	Total
556	1179	1735

3.2. Data Cleaning and Pre-Processing

Data cleaning is vital for sentiment analysis. Tweets that are finalized for analysis contain some raw text which is required to be cleaned and normalized. The data are cleaned by removing hashtags (#), re-tweets (RTs), user handles, multiple spaces, html tags and short words. Moreover, URLs, hyperlinks, special characters, numeric data and emoticons are removed. Stop words (to, on, is, are, am) were removed as these words do not affect text classification. Lemmatization was performed to convert the word to its base form using Natural Language Tool Kit (NLTK) library. POS tagging can be used as key to extract opinion words. We used NLTK POS tagger to extract sentiment words from the tweets to create knowledgebase.

Sentiment analysis is a multiclass classification task that can classify a given text into positive, neutral or negative class. Sentiment analysis is an NLP technique that is used to analyze people's attitudes or emotions in a text or review. It helps to know what the people like or dislike. Based on the nature of training dataset, classification can be binary-class classification with positive and negative sentiments only, or it can be one of three classes (positive, negative and neutral) of sentiment. Python comes in with several packages that can be used for different tasks. TextBlob, VADER and Flair are some of the packages that can be used for sentiment analysis. Each of these packages uses different methods to perform sentiment analysis.

3.3. Proposed Hybrid Model for Sentiment Analysis

For this research, hybrid approach was adopted to analyze learners' sentiments regarding online education system during COVID-19 pandemic. Here, the hybrid approach is the combination of methods (machine learning, ensemble) with customized knowledgebase. This knowledgebase is the combination of sentiment lexicons, SentiWordNet [54] and tweet words (sentiments) that were extracted from the dataset by utilizing part-of-speech (POS) tagger (adjectives).

After the customization of knowledgebase, the classification of the text was performed by implementing different classifiers that use supervised approach along with the variations of n-grams (unigram, bigram, trigram and mixed (1,3) gram) on machine learning and k-fold cross-validation models. TF-IDF feature selection method was used for this study. Selection of features in data that contribute to the prediction is important as this reduces overfitting, improves accuracy and reduces training time while removing irrelevant features that would decrease the accuracy of the models. The TF-IDF matrix in this study comprised words that were in tweets and lexicon. Each n-gram is a combination of words that are in the list of lexicons (knowledgebase-1, knowledgebase-2 and tweet words). It was found that for most of the tweet words, the feature set of knowledgebase-2 was same but different in case of some tweets. Because the raw dataset had mixture of both positive and negative tweets and was imbalanced, experiments (using machine learning methods) were performed with SMOTE technique to see the improvements in the results. Following machine learning methods were applied in combination with knowledgebase:

- Support Vector Classifier (SVC);
- Multinomial Naïve Bayes (MNB);
- Gaussian Naïve Bayes (GNB);
- K-nearest neighbors (KNNs);
- Logistic regression (LR);
- Decision tree (DT);
- Random forest (RF);

- AdaBoost (ADB).

In addition to the above-mentioned techniques, an ensemble model that comprised best classification methods with each of the n-gram (SVC, DT, RF, ADB and LR) was used. This proposed model and its parameters were evaluated using Python programming and its various libraries on the dataset that comprised tweets related to the public views on online education during COVID-19. N-gram model implemented various possible combinations of words along with different settings of knowledgebase to improve the efficacy of models. The customization of knowledgebase with its justification is explained in Section 3.3.1.

3.3.1. Customized Knowledgebase of Sentiments

For the customization of knowledgebase, Python's library TextBlob was used to analyze the polarity and subjectivity of a given sentence and assign the polarity to each individual word and phrase. Final sentiment score was calculated by averaging the score of all sentiments. The polarity score "0" denotes a negative sentiment, and "1" denotes a positive sentiment. TextBlob calculates subjectivity by using parameter "intensity". Intensity determines if a word modifies the next word. We created knowledgebase of opinion words based on the sentiment analysis results of TextBlob and SentiWordNet as follows:

1. All positive and negative predicted results of both TextBlob and SentiWordNet were selected in separate files.
2. In next step, synonyms and antonyms of all the positive and negative words were chosen using SentiWordNet.

To check the performance of machine learning and ensemble models, we created three types of lexicons and knowledgebases as tabulated in Table 4. "AllW" is a customized opinion word list that is of different domain and was also used to evaluate the effect of lexicons from other domains.

Table 4. Opinion words generated from tweets and customized knowledgebase.

Settings	Lexicon Type	Number of Words
1.	Tweets words (TW)	3335
2.	Customized Knowledgebase-1 (Tweets words + Synonyms and Antonyms (TWSA))	11,382
3.	Customized Knowledgebase-2 (Tweets words + Synonyms and Antonyms + Customized annotated words (AllW))	17,948

4. Main Results

Imbalanced datasets can be dealt with in two ways: The first is "Undersampling" in which the sample size of the majority class is reduced. The second is "Oversampling" in which the size of the minority class increases. Oversampling increases training time and causes overfitting, whereas undersampling may cause data loss. Term frequency-inverse document frequency (TF-IDF) was adopted as a weighting method to represent the extracted features as a numeric vector. The performance of individual classifiers, as well as the ensemble method, was evaluated with SMOTE (oversampling) using TF-IDF n-grams (unigram, bigram, trigram and (1,3)). All these methods and features employed three different settings of the lexicon and knowledgebase. The results of all these are discussed in the next section.

4.1. Machine Learning Classifiers Applied Using TF-IDF Unigram and Knowledgebase Settings

This classification step applied the classifiers on the TF-IDF unigram of the tweet dataset which was divided into a training set (80%) and a testing set (20%). The selected features comprised the tokens extracted from the tweets. The ensemble (SVC, DT, RF, ADB and LR) approach applied to the unigram with SMOTE (oversampling) outperformed the

other classifiers and acquired 79.18% and 79.62% accuracy. Results of the other classifiers are also mentioned in Table 5.

Table 5. Results of machine learning classifiers with TF-IDF unigrams and tweet words.

Machine Learning Classifiers Used	TF-IDF Unigrams with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold (%)
SVC	75.8	73.59
GNB	72.01	67.74
MNB	73.18	71.03
KNN	36.15	69.15
LR	76.67	67.06
DT	68.22	69.17
RF	73.17	64.71
ADB	68.8	66.82
Ensemble of (SVC, DT, RF, ADB, LR)	79.18	79.62

All the classifiers using TF-IDF unigrams were then experimented with along with the customized knowledgebase-1. The details of the results of all classifiers are given in Table 6. It can be observed that using 5-fold cross-validation, the ensemble method (74.41%) outperformed all the other methods, while the accuracy of the RF (78.13%) and LR (76.77%) also showed promising results.

Table 6. Results of machine learning classifiers using TF-IDF unigrams and customized knowledgebase-1.

Machine Learning Classifiers Used	TF-IDF Unigrams with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold (%)
SVC	74.34	72.66
GNB	62.09	61.67
MNB	74.93	70.1
KNN	67.05	67.76
LR	76.77	71.05
DT	70.85	60.06
RF	78.13	63.75
ADB	72.01	67.28
Ensemble of (SVC, DT, RF, ADB, LR)	74.52	74.41

Next, classifiers were applied with unigrams and lexicons that comprised knowledgebase-2. The ensemble (SVC, DT, RF, ADB and LR) classifier with SMOTE produced 75.13% accuracy. However, in the evaluation using 5-fold cross-validation, the ensemble method outperformed the others with 75.82% accuracy. The performance of the other classifiers is tabulated in Table 7.

4.2. Machine Learning Classifiers Applied Using TF-IDF Bigrams and Knowledgebase Settings

In this experiment, we applied the individual classifiers using TF-IDF bigrams of the dataset. The dataset was divided into training and testing sets with a ratio of 80% to 20%, respectively. The ensemble (SVC, DT, RF, ADB and LR) classifier using TF-IDF bigrams performed well with SMOTE with an accuracy of 73.73% and 5-fold cross-validation of 73.64%. Details of the results are shown in Table 8.

Table 7. Results of machine learning classifiers using TF-IDF unigrams and customized knowledgebase-2.

Machine Learning Classifiers Used	TF-IDF Unigrams with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold (%)
SVC	74.34	73.12
GNB	68.22	71.49
MNB	72.89	69.16
KNN	70.85	69.38
LR	72.89	69.39
DT	73.47	64.25
RF	74.63	63.54
ADB	65.59	69.38
Ensemble of (SVC, DT, RF, ADB, LR)	75.13	75.82

Table 8. Results of machine learning classifiers with TF-IDF bigrams and tweet words.

Machine Learning Classifiers Used	TF-IDF Bigrams with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold (%)
SVC	74.93	71.03
GNB	63.56	55.13
MNB	62.97	68.69
KNN	27.98	62.86
LR	74.93	67.99
DT	58.31	67.52
RF	74.05	61.72
ADB	71.72	67.07
Ensemble of (SVC, DT, RF, ADB, LR)	73.73	73.64

In the next step, TF-IDF bigrams of the dataset were used along with the customized knowledgebase-1. The ensemble (SVC, DT, RF, ADB and LR) classifier using TF-IDF bigrams performed well with SMOTE with an accuracy of 75.04% and 5-fold cross-validation of 75.13%. Details of the results are shown in Table 9.

After the application of classifiers using setting#2 (using customized knowledgebase-1), TF-IDF bigrams of the dataset were then used to evaluate the performance of machine learning methods and ensemble method in combination with customized knowledgebase-2. The accuracy of the ensemble (SVC, DT, RF, ADB and LR) applied on this data configuration was the best compared to all other classifiers; the accuracies were 77.20% and 77.36% with SMOTE (oversampling). Detailed results are defined in Table 10.

4.3. Machine Learning Classifiers Applied Using TF-IDF Trigrams and Knowledgebase Settings

In this experiment, we applied the individual classifiers and ensemble method using TF-IDF trigrams of the dataset. In comparison with conventional approaches, the ensemble classifier yielded 74.34% accuracy with 5-folds and with SMOTE oversampling. Complete results obtained by all other classifiers are shown in Table 11.

Table 9. Results of machine learning classifiers with TF-IDF bigrams and customized knowledgebase-1.

Machine Learning Classifiers Used	TF-IDF Bigrams with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold (%)
SVC	70.55	70.09
GNB	53.35	40.89
MNB	51.6	67.06
KNN	60.93	62.59
LR	70.85	71.26
DT	72.88	68.92
RF	66.76	68.22
ADB	65.31	69.39
Ensemble of (SVC, DT, RF, ADB, LR)	75.04	75.13

Table 10. Results of machine learning classifiers with bigram and customized knowledgebase-2.

Machine Learning Classifiers Used	TF-IDF Bigram with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold (%)
SVC	73.76	72.66
GNB	63.27	52.09
MNB	64.72	69.40
KNN	33.82	61.71
LR	74.93	67.76
DT	72.59	63.78
RF	69.09	64.96
ADB	69.97	71.96
Ensemble of (SVC, DT, RF, ADB, LR)	77.20	77.36

Table 11. Results of machine learning classifiers with TF-IDF trigrams and tweet words.

Machine Learning Classifiers Used	TF-IDF Trigrams with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold (%)
SVC	73.76	68.7
GNB	50.15	39.47
MNB	49.56	68.45
KNN	61.8	59.32
LR	73.18	69.62
DT	47.81	64.52
RF	66.76	56.14
ADB	67.05	67.06
Ensemble of (SVC, DT, RF, ADB, LR)	74.31	74.34

Similarly, the classifiers were applied using TF-IDF trigrams of the dataset and with knowledgebase-1. It can be observed that individual classifiers did not perform well with TF-IDF trigrams. However, LR, RF, ADB and DT contributed better scores, and the

ensemble method scored 74.02% accuracy and 74.05% applying 5-fold cross-validation. Details of the results with all other classifiers are given in Table 12.

Table 12. Results of machine learning classifiers with trigram and knowledgebase-1.

Machine Learning Classifiers Used	TF-IDF Trigram with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold(%)
SVC	44.31	67.99
GNB	45.48	35.04
MNB	45.48	66.82
KNN	67.05	67.06
LR	74.05	68.45
DT	73.47	70.79
RF	68.22	68.92
ADB	68.51	67.52
Ensemble of (SVC, DT, RF, ADB, LR)	74.02	74.05

Similar to the previous experiment, individual classifiers and the ensemble method were applied using TF-IDF trigrams of the dataset along with knowledgebase-2. The ensemble classifier produced 76.32% accuracy and 75.21% with 5-fold cross-validation, which was the highest accuracy against the conventional machine learning classifiers. Detailed results of all machine learning classifiers are represented in Table 13.

Table 13. Results of machine learning classifiers with trigram with knowledgebase-2.

Machine Learning Classifiers Used	TF-IDF Trigram with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold (%)
SVC	74.93	71.26
GNB	49.56	38.79
MNB	49.85	68.22
KNN	72.01	57.74
LR	76.09	67.99
DT	76.09	71.96
RF	69.68	62.34
ADB	70.26	67.29
Ensemble of (SVC, DT, RF, ADB, LR)	76.32	75.21

4.4. Machine Learning Classifiers Applied Using (1,3) Gram and Knowledgebase Settings

The classifiers were evaluated with a mixed gram (a set of one, two and three grams). The dataset was divided into 80% of training and 20% of testing data. The ensemble (SVC, DT, RF, ADB and LR) algorithm obtained 78.71% accuracy and 79.30% accuracy with 5-fold cross-validation using SMOTE (oversampling). With this combination, the logistic regression classifier also performed well with an accuracy of 79.01%. The selected lexicon for this comprised tweet words only. Complete results with this setting are tabulated in Table 14.

Table 14. Results of machine learning classifiers with (1,3) gram and tweet words.

Machine Learning Classifiers Used	TF-IDF (1,3) Gram with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold (%)
SVC	76.17	67.99
GNB	74.34	70.12
MNB	75.8	70.33
KNN	33.24	61.45
LR	79.01	67.29
DT	71.43	67.75
RF	72.88	69.39
ADB	72.01	71.97
Ensemble of (SVC, DT, RF, ADB, LR)	78.71	79.30

The next step of (1,3) gram was implemented with knowledgebase-1. It was observed that this time the SVC classifier and ensemble model yielded results with 76.06% and 75.80% accuracy. Details of the results are tabulated in Table 15.

Table 15. Results of machine learning classifiers with (1,3) gram and knowledgebase-1.

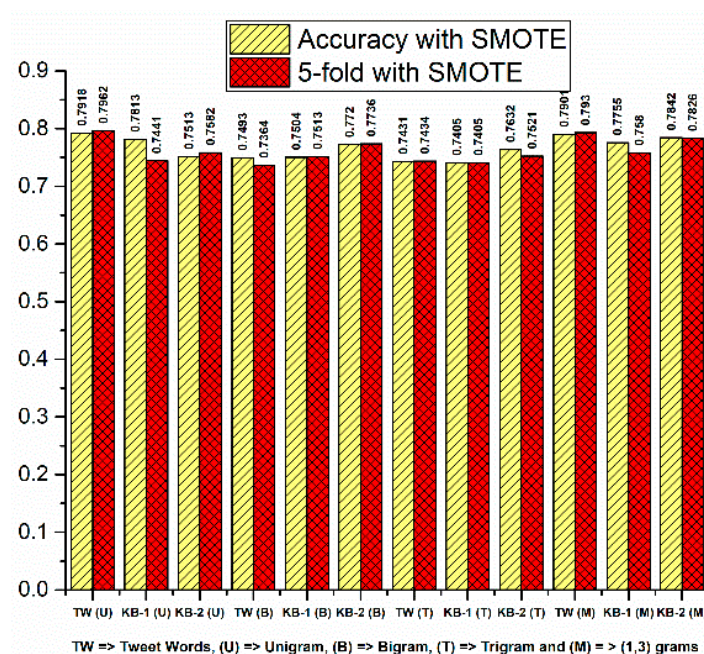
Machine Learning Classifiers Used	TF-IDF (1,3) Gram with SMOTE	
	Accuracy (%)	Cross-Validation with 5-Fold (%)
SVC	74.64	72.42
GNB	66.76	64.49
MNB	74.63	69.16
KNN	68.51	71.51
LR	77.55	70.32
DT	72.01	67.76
RF	76.09	67.07
ADB	72.3	69.17
Ensemble of (SVC, DT, RF, ADB, LR)	76.06	75.80

In the last step, the classifiers and ensemble method with (1,3) gram of the dataset including knowledgebase-2 were used. The ensemble classifier performed the best for imbalanced data with an accuracy of 78.42% and with 5-fold of 78.26%. Results of all other classifiers are formulated in Table 16.

As per the observation of all the settings and n-gram features (unigram, bigram, trigram, combination (1,3)), the overall performance of the ensemble classifier with SMOTE was evaluated. It can be seen that in the case of using SMOTE to tackle the data imbalance issue, our hybrid method that incorporated the ensemble classifier with SMOTE performed better with the unigram feature and tweet words (accuracy = 79.18%, 5-fold cross-validation = 79.62%) as evident from Figure 1. Moreover, the ensemble classifier also obtained better results in the case of using TF-IDF (1,3) grams and knowledgebase-2 (accuracy = 78.42%, 5-fold cross-validation = 78.26%).

Table 16. Results of machine learning classifiers with TF-IDF (1,3) grams and knowledgebase-2.

Machine Learning Classifiers Used	TF-IDF (1,3) Grams with SMOTE	
	Accuracy (%)	Cross-Validation with 5-fold (%)
SVC	72.01	71.73
GNB	76.38	74.76
MNB	78.13	68.46
KNN	36.73	69.85
LR	76.97	67.98
DT	72.59	69.88
RF	72.59	70.32
ADB	73.76	69.39
Ensemble of (SVC, DT, RF, ADB, LR)	78.42	78.26

**Figure 1.** Performance measurements of hybrid method with SMOTE (ensemble method using n-grams and all settings of knowledgebase).

5. Conclusions

In this research, the effectiveness of the online education system during the COVID-19 pandemic was evaluated by analyzing learner reviews that were collected (from the year 2020) from Twitter using relevant hashtags. Using the proposed automatic analysis of learners' sentiments, the performance of different machine learning classifiers in combination with the n-gram technique was analyzed with different settings of the knowledgebase (tweet words, knowledgebase-1 and knowledgebase-2). In addition to this, the ensemble method (with unigram and tweet words) was applied and found to be effective against conventional machine learning classifiers on this dataset. Moreover, due to imbalance in the nature of the collected dataset, SMOTE (oversampling) was used in combination with both of the mentioned approaches (machine learning classifiers and ensemble method). It was observed that the performance of the hybrid approach incorporating the ensemble model is improved with most of the settings as compared to individual machine learning classifiers. It was also noted that the performance of the machine learning and ensemble classifiers is improved with the customization of the lexicon to knowledgebase-1 and knowledgebase-2.

Studies have used deep learning techniques but without customization of a domain-specific knowledgebase. Future work of this research should involve collection of more data [55–57] to use deep learning models with a customized (domain-specific) knowledgebase.

Overall, the sentiments were found to be positive indicating that learners generally adapted well to the online mode of education during the COVID-19 pandemic. At the same time, some crucial aspects were also identified as a result of the sentiment analysis of the acquired data. Some of the more important aspects pertaining to the online mode of education were “lack of resources” and “health issues”. Thus, this study provides a basic framework for automatic sentiment analysis leading to identification of important aspects that have a significant role to play in ensuring the effectiveness and long-term sustainability of the online mode of education.

Author Contributions: Conceptualization and Data Analysis, M.I., S.H. and M.M.B.; Data Cleaning and Methodology, M.I.; Formal analysis and review, S.H.; Supervision, S.H. and M.M.B.; Writing—Original Draft Preparation, M.I.; Writing—Reviewing and Editing, S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NED University of Engineering & Technology research funds.

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.

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

References

1. Tiffani, I.E. Optimization of Naïve Bayes Classifier by Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review. *J. Soft Comput. Explor.* **2020**, *1*, 1–7. [CrossRef]
2. Al-Hashedi, A.; Al-Fuhaidi, B.; Mohsen, A.M.; Ali, Y.; Al-Kaf, H.A.G.; Al-Sorori, W.; Maqtary, N. Ensemble Classifiers for Arabic Sentiment Analysis of Social Network (Twitter Data) towards COVID-19-Related Conspiracy Theories. *Appl. Comput. Intell. Soft Comput.* **2022**, *2022*, 6614730. [CrossRef]
3. Wolff, L.-A. Sustainability Education in Risks and Crises: Lessons from COVID-19. *Sustainability* **2020**, *12*, 5205. [CrossRef]
4. Li, C.; Zhou, H. Enhancing the Efficiency of Massive Online Learning by Integrating Intelligent Analysis into MOOCs with an Application to Education of Sustainability. *Sustainability* **2018**, *10*, 468. [CrossRef]
5. Faura-Martínez, U.; Lafuente-Lechuga, M.; Cifuentes-Faura, J. Sustainability of the Spanish university system during the pandemic caused by COVID-19. *Educ. Rev.* **2021**, 1–19. [CrossRef]
6. Ionescu, C.A.; Paschia, L.; Gudanescu Nicolau, N.L.; Stanescu, S.G.; Neacsu Stancescu, V.M.; Coman, M.D.; Uzlau, M.C. Sustainability Analysis of the E-Learning Education System during Pandemic Period—COVID-19 in Romania. *Sustainability* **2020**, *12*, 9030. [CrossRef]
7. Rehman, A.U. Challenges to Online Education in Pakistan During COVID-19 & the Way Forward. 2020. Available online: <https://preprints.aijr.org/index.php/ap/preprint/view/241> (accessed on 27 January 2022).
8. Yang, R. Machine Learning and Deep Learning for Sentiment Analysis Over Students’ Reviews: An Overview Study. 2021. Available online: <https://www.preprints.org/manuscript/202102.0108/v1> (accessed on 27 January 2022).
9. Manguri, K.H.; Ramadhan, R.N.; Amin, P.R. Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks. *Kurd. J. Appl. Res.* **2020**, *5*, 54–65. [CrossRef]
10. Wang, Y.; Chen, Q.; Shen, J.; Hou, B.; Ahmed, M.; Li, Z. Aspect-level sentiment analysis based on gradual machine learning. *Knowl.-Based Syst.* **2020**, *212*, 106509. [CrossRef]
11. Zhao, P.; Hou, L.; Wu, O. Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *Knowl.-Based Syst.* **2019**, *193*, 105443. [CrossRef]
12. Pota, M.; Ventura, M.; Catelli, R.; Esposito, M. An Effective BERT-Based Pipeline for Twitter Sentiment Analysis: A Case Study in Italian. *Sensors* **2021**, *21*, 133. [CrossRef]
13. Sindhu, I.; Daudpota, S.M.; Badar, K.; Bakhtyar, M.; Baber, J.; Nurunnabi, M. Aspect-Based Opinion Mining on Student’s Feedback for Faculty Teaching Performance Evaluation. *IEEE Access* **2019**, *7*, 108729–108741. [CrossRef]
14. Li, N.; Chow, C.-Y.; Zhang, J.-D. SEML: A Semi-Supervised Multi-Task Learning Framework for Aspect-Based Sentiment Analysis. *IEEE Access* **2020**, *8*, 189287. [CrossRef]
15. Nikolić, N.; Grljević, O.; Kovačević, A. Aspect-based sentiment analysis of reviews in the domain of higher education. *Electron. Libr.* **2020**, *38*, 44–64. [CrossRef]

16. Nandal, N.; Tanwar, R.; Pruthi, J. Machine learning based aspect level sentiment analysis for Amazon products. *Spat. Inf. Res.* **2020**, *28*, 601–607. [CrossRef]
17. Mowlaei, M.E.; Abadeh, M.S.; Keshavarz, H. Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Syst. Appl.* **2020**, *148*, 113234. [CrossRef]
18. Liu, N.; Shen, B. Aspect-based sentiment analysis with gated alternate neural network. *Knowl.-Based Syst.* **2019**, *188*, 105010. [CrossRef]
19. Feng, J.; Gong, C.; Li, X.; Lau, R.Y.K. Automatic Approach of Sentiment Lexicon Generation for Mobile Shopping Reviews. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 9839432. [CrossRef]
20. Lee, G.; Jeong, J.; Seo, S.; Kim, C.; Kang, P. Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. *Knowl.-Based Syst.* **2018**, *152*, 70–82. [CrossRef]
21. Ma, Y.; Peng, H.; Khan, T.; Cambria, E.; Hussain, A. Sentic LSTM: A Hybrid Network for Targeted Aspect-Based Sentiment Analysis. *Cogn. Comput.* **2018**, *10*, 639–650. [CrossRef]
22. Amrani, Y.A.; Lazaar, M.; El Kadiri, K.E. Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. *Procedia Comput. Sci.* **2018**, *127*, 511–520. [CrossRef]
23. Hazarika, D.; Konwar, G.; Deb, S.; Bora, D.J. Sentiment Analysis on Twitter by Using TextBlob for Natural Language Processing. *ICRMAT* **2020**, *24*, 63–67. [CrossRef]
24. Imran, M.; Mitra, P.; Carlos, C. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. Available online: <https://arxiv.org/abs/1605.05894> (accessed on 27 January 2022).
25. Ahuja, R.; Chug, A.; Kohli, S.; Gupta, S.; Ahuja, P. The Impact of Features Extraction on the Sentiment Analysis. *Procedia Comput. Sci.* **2019**, *152*, 341–348. [CrossRef]
26. Srividya, K.; Sowjanya, A.M. Aspect Based Sentiment Analysis using POS Tagging and TFIDF. *Int. J. Eng. Adv. Technol. IJEAT* **2019**, *8*, 1960–1963. [CrossRef]
27. Yang, Y. Research and Realization of Internet Public Opinion Analysis Based on Improved TF-IDF Algorithm. In Proceedings of the International Symposium on Distributed Computing and Applications to Business, Engineering and Science, Anyang, China, 13–16 October 2017.
28. Li, Y.; Shen, B. Research on Sentiment Analysis of Microblogging Based on LSA and TF-IDF. In Proceedings of the 3rd IEEE International Conference on Computer and Communications, Chengdu, China, 13–16 December 2017.
29. Esuli, A.; Sebastiani, F. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Available online: <https://aclanthology.org/L06-1225/> (accessed on 27 January 2022).
30. Guerini, M.; Gatti, L.; Turchi, M. Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet. *arXiv* **2013**, arXiv:1309.5843. [CrossRef]
31. Asghar, D.M. Detection and Scoring of Internet Slangs for Sentiment Analysis Using SentiWordNet. *Life Sci. J.* **2014**, *11*, 66–72.
32. Khan, F.H.; Qamar, U.; Bashir, S. SWIMS: Semi-supervised subjective feature weighting and intelligent model selection for sentiment analysis. *Knowl.-Based Syst.* **2016**, *100*, 97–111. [CrossRef]
33. Tierney, B. Sentiment Classification of Reviews Using SentiWordNet. In Proceedings of the 9th IT&T Conference, Technological University Dublin, Dublin, Ireland, 22–23 October 2009.
34. Husnain, M.; Missen, M.M.; Akhtar, N.; Coustaty, M.; Mumtaz, S.; Prasath, V.B. A systematic study on the role of SentiWordNet in opinion mining. *Front. Comput. Sci.* **2019**, *15*, 154614. [CrossRef]
35. Kastrati, Z.; Imran, A.S.; Kurti, A. Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs. *IEEE Access* **2020**, *8*, 106799–106810. [CrossRef]
36. Ishaq, A.; Asghar, S.; Gillani, S.A. Aspect-Based Sentiment Analysis Using a Hybridized Approach Based on CNN and GA. *IEEE Access* **2020**, *8*, 135499–135512. [CrossRef]
37. Da'U, A.; Salim, N.; Rabi, I.; Osman, A. Recommendation system exploiting aspect-based opinion mining with deep learning method. *Inf. Sci.* **2019**, *512*, 1279–1292. [CrossRef]
38. Madhoushi, Z.; Hamdan, A.R.; Zainudin, S. Aspect-Based Sentiment Analysis Methods in Recent Years. *Asia-Pac. J. Inf. Technol. Multimed.* **2019**, *08*, 79–96.
39. Yang, T.; Yin, Q.; Yang, L.; Wu, O. Aspect-based Sentiment Analysis with New Target Representation and Dependency Attention. *IEEE Trans. Affect. Comput.* **2019**, *1*. [CrossRef]
40. Xing, Y.; Xiao, C.; Wu, Y.; Ding, Z. A Convolutional Neural Network for Aspect-Level Sentiment Classification. *Int. J. Pattern Recognit. Artif. Intell.* **2019**, *33*, 1959046. [CrossRef]
41. Nandhini, M.D.S.; Pradeep, G. A Hybrid Co-occurrence and Ranking-based Approach for Detection of Implicit Aspects in Aspect-Based Sentiment Analysis. *SN Comput. Sci.* **2020**, *1*, 128. [CrossRef]
42. Araque, O.; Corcuera-Platas, I.; Sánchez-Rada, J.F.; Iglesias, C.A. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Syst. Appl.* **2017**, *77*, 236–246. [CrossRef]
43. Pham, D.H.; Nguyen, T.T.; Le, A.C. *Fine-Tuning Word Embeddings for Aspect-Based Sentiment Analysis*; Springer: Cham, Switzerland, 2017; pp. 500–508.
44. Zainuddin, N.; Selamat, A.; Ibrahim, R. Hybrid sentiment classification on twitter aspect-based sentiment analysis. *Appl. Intell.* **2017**, *48*, 1218–1232. [CrossRef]

45. Pham, D.-H.; Le, A.-C. Learning multiple layers of knowledge representation for aspect based sentiment analysis. *Data Knowl. Eng.* **2018**, *114*, 26–39. [[CrossRef](#)]
46. Pham, D.-H.; Le, A.-C. Exploiting multiple word embeddings and one-hot character vectors for aspect-based sentiment analysis. *Int. J. Approx. Reason.* **2018**, *103*, 1–10. [[CrossRef](#)]
47. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning Based Text Classification: A Comprehensive Review. *ACM Comput. Surv.* **2021**, *54*, 1–40. [[CrossRef](#)]
48. Vizcarra, J.; Kozaki, K.; Ruiz, M.T.; Quintero, R. Knowledge-Based Sentiment Analysis and Visualization on Social Networks. *New Gener. Comput.* **2020**, *39*, 199–229. [[CrossRef](#)]
49. Van Atteveldt, W.; van der Velden, M.A.; Boukes, M. The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Commun. Methods Meas.* **2021**, *15*, 121–140. [[CrossRef](#)]
50. Bayerl, P.S.; Paul, K.I. What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. *Comput. Linguist.* **2011**, *37*, 699–725. [[CrossRef](#)]
51. Bhowmick, P.K.; Basu, A.; Mitra, P. An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Text. In Proceedings of the Workshop on Human Judgements in Computational Linguistics, Manchester, UK, 23 August 2008; pp. 58–65.
52. Goldberg, D.M.; Khan, S.; Zaman, N.; Gruss, R.J.; Abrahams, A.S. Text Mining Approaches for Postmarket Food Safety Surveillance Using Online Media. *Risk Anal.* **2020**. [[CrossRef](#)] [[PubMed](#)]
53. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [[CrossRef](#)] [[PubMed](#)]
54. Baccianella, S.; Esuli, A.; Sebastiani, F. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), Valletta, Malta, 19–21 May 2010.
55. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R.; et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* **2021**, *76*, 243–297. [[CrossRef](#)]
56. Feng, S.; Zhou, H.; Dong, H. Using deep neural network with small dataset to predict material defects. *Mater. Des.* **2018**, *162*, 300–310. [[CrossRef](#)]
57. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.