IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 24 August 2022, accepted 12 September 2022, date of publication 22 September 2022, date of current version 4 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3208953

RESEARCH ARTICLE

Data Analytics on Online Student Engagement Data for Academic Performance Modeling

XIAOHUI TAO[1], (Senior Member, IEEE), AARON SHANNON-HONSON[1], PATRICK DELANEY[1], LIN LI[2], (Member, IEEE), CHRISTOPHER DANN[3], YAN LI[1], AND HAORAN XIE[4], (Senior Member, IEEE)

[1]School of Mathematics, Physics and Computing, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[2]School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430062, China
[3]School of Education, University of Southern Queensland, Springfield, QLD 4300, Australia
[4]Department of Computing and Decision Sciences, Lingnan University, Hong Kong, SAR

Corresponding author: Xiaohui Tao (xiaohui.tao@usq.edu.au)

This work was supported in part by the Academic Transformation Portfolio Unit in the University of Southern Queensland.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Human Research Ethics Committee from the University of Southern Queensland under Approval No. H20REA137.

ABSTRACT In large MOOC cohorts, the sheer variance and volume of discussion forum posts can make it difficult for instructors to distinguish nuanced emotion in students, such as engagement levels or stress, purely from textual data. Sentiment analysis has been used to build student behavioral models to understand emotion, however, more recent research suggests that separating sentiment and stress into different measures could improve approaches. Detecting stress in a MOOC corpus is challenging as students may use language that does not conform to standard definitions, but new techniques like TensiStrength provide more nuanced measures of stress by considering it as a spectrum. In this work, we introduce an ensemble method that extracts feature categories of engagement, semantics and sentiment from an AdelaideX student dataset. Stacked and voting methods are used to compare performance measures on how accurately these features can predict student grades. The stacked method performed best across all measures, with our Random Forest baseline further demonstrating that negative sentiment and stress had little impact on academic results. As a secondary analysis, we explored whether stress among student posts increased in 2020 compared to 2019 due to COVID-19, but found no significant change. Importantly, our model indicates that there may be a relationship between features, which warrants future research.

INDEX TERMS Ensemble method, natural language processing, MOOC, academic performance modeling.

I. INTRODUCTION

Discussion forums provide a crucial point of contact for Massively Online Open Courses (MOOCs), where instructors and students communicate about course content, assignment queries and general socialisation. However, the large numbers of participants and the sheer variance and volume of posts can make it difficult for instructors to gain a sense of the emotional state of their cohort, which may be important in student outcomes. This has motivated studies such as [1] and [2] to develop approaches to detect salient features among forum 'noise' and provide ways for instructors to identify urgent posts for timely intervention. Many studies have used sentiment analysis to interpret student behavior in MOOC courses through discussion forum posts [3], [4], [5], [6]. While sentiment is useful for understanding opinions, attitudes and emotion, more recent studies have sought to distinguish further nuances in features such as stress to develop more holistic models of student behavior. The challenge of detecting sentiment and stress in a MOOC corpus is that language used by students may not always conform to standard meanings. Therefore, detecting stress requires refined methods, as demonstrated by the development of models such as TensiStrength [7].

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu.

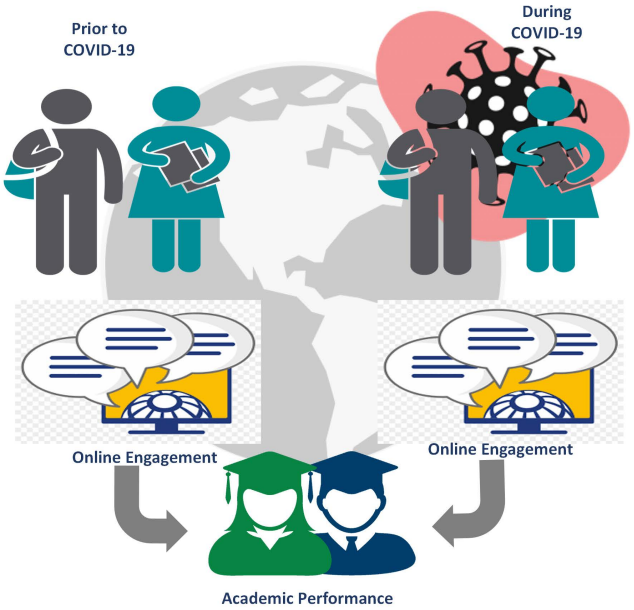


FIGURE 1. MOOC student performance modelling by online engagement analysis.

Students studying through a wholly online mode are likely already predisposed to a number of life stressors, such as family commitments, full-time work or living in remote or isolated regions. The effects of COVID-19 in 2020 on brick and mortar tertiary students has been widely reported, particularly with its impact on mental health [8], [9], but there is still a need to investigate the impact of the pandemic on MOOC students. In particular, a comparison between pre- and during-COVID behaviors will give insight into any behavioral changes resulting from this event, allowing researchers to understand how it impacted sentiment and stress in online learning modes. Our study uses an innovative approach in which we separate sentiment and stress into more nuanced representations, perceiving them as a spectrum rather than binary values, similar to work by [10].

In this work, we use 2019 and 2020 data from the AdelaideX platform to learn a student behavioral model combining features such as engagement and semantics with sentiment and stress measures. Our research objective is to use these features to predict academic performances using an ensemble model, specifically to (1) compare the differences in these features between a pre and during-COVID student cohort; (2) observe the impact of stress on student academic performance between these years; and (3) rank the features in terms their overall importance to student outcomes. Figure 1 illustrates the concept of the study, which compares students' online engagement prior to and during COVID-19 to gain in-depth understanding of features in student behavior that impact academic performance. We are motivated by the need to refine existing techniques that extract meaningful representations of student behavior from textual data. TensiStrength has previously been used to analyse short-form social media texts, with a specific concentration on stress related to public transport [7]. We apply TensiStrength to the MOOC domain and test its applicability using student posting data, which may also be subject to event-specific stressors.

The work is driven by the following research questions:

1) How does stress compare to other discussion forum features such as engagement, semantic and sentiment in determining student academic performance?
2) Did stress increase among student cohorts during the pandemic?

To achieve this, we use an ensemble method consisting of three machine learning algorithms (Naîve Bayes, Random Forests and Deep Learning), with overall results filtered using stacked and voting methods. TensiStrength is used to extract stress features and provide numeric calculations for sentiment and stress measures. This will provide a more measured understanding of the impact of COVID-19 on online learning. As far as we are aware, this study is one of the first to utilise TensiStrength in the educational space for detecting stress. Our overarching contributions are the following:

- A method for distinguishing student engagement, semantics and emotional measures such as sentiment and stress
- An approach for ranking these features in terms of their importance on student academic performance
- A model of student behavior built on granular feature extraction to give MOOC instructors more insight into the emotional state within student posts
- The development of an ensemble model that uses multiple algorithms to produce the most accurate results

The development of the ensemble model provides a platform-agnostic tool that can assist in identifying posts that require urgent intervention, adding both theoretical and methodological contributions to the MOOC research domain.

The paper is structured as follows. After this Introduction, Section II discusses state-of-the-art works related to our study. The research problem and aim are defined in Section III, followed by the research design and technical details presented in Section IV. In Section V, the experiment design and experimental results are reported, with the related discussions presented in Section VI. Finally, Section VII provides the conclusion and discusses future work.

II. RELATED WORK

Data-mining techniques are well-established in Social Media research for retrieving textual content to model user behavior [11], [12], [13], [14], [15]. While sentiment studies have made significant advances into health fields such as mental health (e.g., [16], [17], the application of machine learning techniques to educational settings such as MOOCs is still developing. A study by [18] determined that standard sentiment analysis methods such as those used in social media research were unsuitable for the MOOC context. Instead, they developed a BERT-based sentiment analyzer that outperformed state-of-the-art social media sentiment predictors with 0.94 accuracy. This demonstrates the need for purposed models.

Previously, sentiment analysis has been used for post-course reporting to improve course quality through opinion mining [19]. For these methods to have utility for real-time monitoring of MOOC forums, more nuanced NLP techniques must be employed to build more suitable models of student user behavior [3]. A novel classification method by [20] explored modified key emotional indicators to more accurately determine sentiment among uncommon ways of representing words (e.g.,"*this is bad*" versus "*this is baaaaad*"). Similarly, research by [21] showed that humour, sarcasm, idioms, and irony are often presented as positive sentiment when the intended message may be the opposite.

While studies such as [22] have sought to address emotional engagement levels in students, there is still ongoing research on detecting stress and using it to build emotional models from textual data. For example, TensiStrength [7] measures stress and relaxation as separate metrics. This more distinguished approach may help interpret nuances in student posts more accurately, particularly when there are uncommon expressions and alternative modes of phrasing to denote their feelings about a course. In analysing event-based language, [23], [24] note that event-based posts such as 'running late' may be correlated as stressful language, when in fact this might just be a neutral statement. In [23], they added pre-processing Word Sense Disambiguation (WSD) to the TensiStrength tool, which improved accuracy and performance for defining ambiguous terms in a sentence that had emotional meaning opposite to its literal definition. TensiStrength has previously been applied to detecting stress in tweets pertaining to transport and stressful events within commuter experience [25]. The potential value of TensiStrength is to underpin tools for timely diagnosis of issues from users, highlighted by its direct benefits in industries reliant on customer satisfaction [25]. We can perceive student attitudes on MOOC forums in a similar fashion, by using posts as a representation of the level of satisfaction a user feels at a point in time in the course. Our study presents one of the first attempts to use TensiStrength in the educational domain.

While [26] and [27] indicate that sentiment and emotional models complement one another, in the MOOC space students may post under stressful circumstances, meaning sentiment alone may not be sufficient to build models of user behavior. Wei *et al.* [28] used text classification to identify features such as confusion and urgency alongside sentiment, to enhance insight into real-time behavior and understand which students required urgent intervention. Specific events or reactions to these events might impact opinions, sentiment and mental health states of users [21], [29]. Research by [14] measured temporal factors of sentiment changes such as post density, frequency, and content-oriented posts, while [4] included continuous variables such as message length, positive or negative orientation and the number of responses to analyse textual behavior across the duration of a course. For a more comprehensive model of student behavior, it is necessary to incorporate features beyond sentiment. To this end, [30] analyzed 'burstiness' (posting frequency) at particular temporal points in a course, which may be explain why students demonstrate particular sentiment or stress at different milestones or times in a semester.

In systematic reviews of sentiment analysis in the education domain, [6] and [31] found that Naïve Bayes and Deep Learning were some of the more common techniques used. Similarly, [4] demonstrate that Random Forest is widely used to analyse forum messages. Therefore, we adopt these three algorithms as baseline classifiers for our experiment design over others. Our work combines prior research into the development of a user behavioral model in the MOOC, taking the COVID-19 pandemic as an overarching event. Numerous reports indicate that this event impacted traditional student cohorts, but there is presently a lack of understanding about its effect on MOOC students. By extracting features such as interaction patterns, common semantic behavior and more nuanced analysis of sentiment and stress, a more holistic model of student behavior in an online context can be determined.

III. RESEARCH AIM

We posit that student behavior can be represented as a set of behavioral features. These features, denoted by f, are quantified or calculated and make up feature vector sets, denoted by FV, which contain each feature weighting. Equation 1 defines these feature vectors.

$$FV = \{\langle f_1, w_1\rangle, \langle f_2, w_2\rangle, \ldots, \langle f_n, w_n\rangle\} \tag{1}$$

Here we use a student's academic performance, defined as *ap*, inside MOOCs as a label for these feature sets. We aim to understand and clarify the coefficients within each behavior set and their impact on both the overall behavior model and *ap* by solving the function $f(\dot{})$ defined by Eq. 2. The coefficients $\alpha, \beta, \ldots, \gamma$ define the impact or importance of each individual feature vector on AP for the model.

$$f(\{\alpha \times \langle f_1, w_1\rangle, \beta \times \langle f_2, w_2\rangle, \ldots \gamma \times \langle f_n, w_n\rangle\}) \rightarrow ap \tag{2}$$

IV. METHODOLOGY

Our conceptual model is depicted in Figure 2. This framework learns the function in Eq. 2 from a data source (solid-lined boxes) and performs data engineering to synthesise additional values (dashed boxes) to finalise the proposed model. Based on observations on MOOC student data, the model is designed to learn from three types of features categories: engagement, semantics and sentiment, as depicted in the Venn Diagram in Figure 3. The learned model is then trained and validated using machine learning prediction algorithms, which outputs a usable instance of our proposed model.

The feature extraction layer is comprised of the feature categories shown in Figure 3. Engagement describes the intensity, or level of interaction, a student has with the discussion forum and incorporates measures determined by overall course activity. We choose to use the total number of *active days*, denoted by *ad*, recorded per student in a course as our temporal measure. An active day in this context is defined as a day where a student has interacted with course content beyond viewing a page, and is one of the measures that does

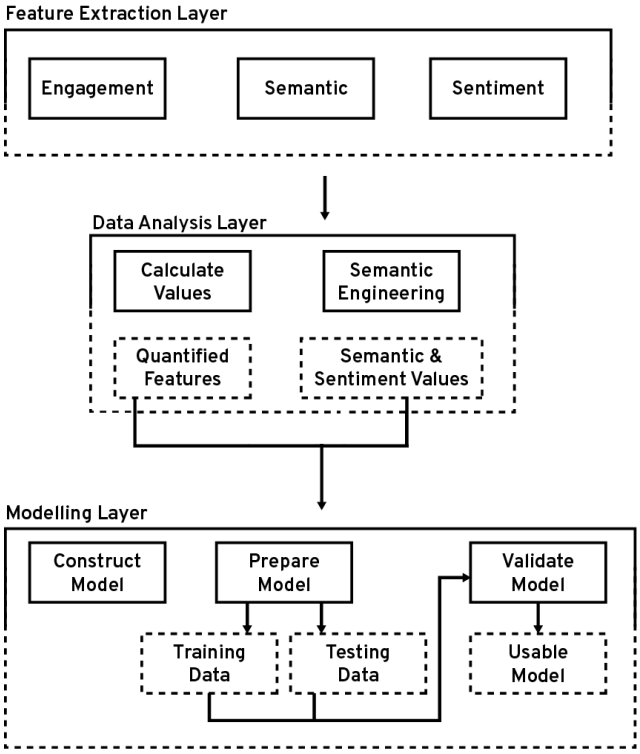


FIGURE 2. Conceptual model for constructing a student-behavior model in MOOCs.

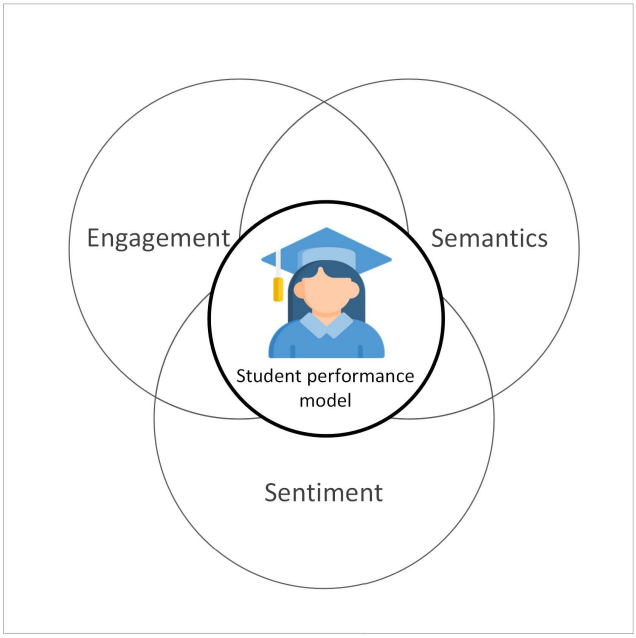


FIGURE 3. Venn diagram of student performance model learned from engagement, semantics, and sentiment features.

not focus purely on the discussion forum. We use the student's post body text (s) to generate a set of semantic features, denoted by SM. An 'on-topic' score, ot is created, which defines how relevant or valuable a post is to the discussion and - by implication - the main topic of the course. ot measures the semantic distance between a post and the description of the correlating course. For example, a post discussing Hamlet made in a MOOC about Shakespeare would have a high ot value, while the same post in a Data Science course would have a low value. A summary of engagement and semantic features extracted at this layer is shown in Table 1.

TABLE 1. Table of engagement and semantic features.

Feature Name	Notation	Description
Number of posts	p	The total number of posts made by a student in the overall course
Post length	l	The length of any individual post, measured in number of words
Active days	ad	The number of days in a course a student is 'active'
Posting ratio	r	A ratio of posts-per-active-day
Study mode	md	A student's mode of study, here defined as students who are paying for a course or completing a course for free
On-topic score	ot	A semantic score of how close the current post is to the overall description of the course or how on topic the current post is

Academic Performance, ap, is used both for self-evaluation for students and a label. From the dataset, ap is originally provided as raw floating point number values. For our purpose, we convert these to an adapted grade scheme reflecting grade milestones at The University of Adelaide [32]: High Distinction (HD) is between 85% and 100%; Distinction (D) is between 75% and 84%; Credit (C) is Between 65% and 74%; Pass (P) is between 50% and 65%; and Fail (F) is 49% and under.

In addition to these engagement and semantic measures, we extract a set of sentiment and stress features denoted by sen and str, respectively. These represent measures of sentiment (how positive or negative a person feels about the topic they are discussing) and stress (how stressed a person feels about the topic they are discussing) among student posts. These measures are often used separately in social media environments [21], [33], [34], but rather than representing them as single spectrum values, here these features are made up of a score describing the intensity of either end of relax/stress or positive/negative sentiment spectrums. Using two separate measures rather than a binary spectrum is valuable as it allows for an element of nuance in sentiment measuring or "mixed feelings" from users. Additionally, we believe that the opposite of stress is not necessarily 'relaxation', but rather a *related* measure, as not all stress is inherently bad. For example, a student with high stress and positive sentiment scores may be more accurately described as 'excited' about something compared to a student with high stress and negative sentiment scores.

TensiStrength uses a lexical approach with manually derived lists of terms related to stress and relaxation [7]. The approach looks to rank terms numerically based on their contextual use, for example, as responses toward situations or states. Differentiating between 'good stress' and 'bad stress' is a valuable addition to our user behavior model. These allow for description of mental state measures and how they impact

TABLE 2. Table of sentiment features extracted.

Feature Name	Notation	Description
Positive sentiment	$+sen$	A measure of 1 to 5 representing the positive sentiment calculated from post content
Negative sentiment	$-sen$	A measure of -5 to -1 representing negative sentiment in the post content
Relaxation	$+str$	A measure of 1 to 5 representing how relaxed a student is from the post content, a positive measure of stress
Stress	$-str$	A measure of -5 to -1 representing how stressed a student is from the post content. Opposite but complementary to relaxation

academic performance as well as the entirety of student behavior. These features are summarised in Table 2.

A. DATASET

The project dataset is sourced from AdelaideX's [1] courses on the EdX platform. This data is solely the property of The University of Adelaide and AdelaideX. For our research scope, we restrict our selections between 1 January and 30 June for the years 2019 and 2020, and use only student-created data. During these periods, the same courses were being offered either as a new course or as a self-paced archived version, so we have similar student cohorts for each year. The 2020 data captures the peak of the pandemic. We use 2019 to initially build the model, and we apply the same methods to 2020 data as a means of testing the model on another dataset. In this initial dataset, there were 4553 students in the 2019 cohort with 6436 posts, while in 2020 there were 4258 total students with 8394 posts. These figures combine both auditing and verified students, which will be separated in our model.

B. FEATURE EXTRACTION

For our approach, we extract the aforementioned features identified from the processed AdelaideX dataset. Some of the features are available in data analytics packages, but the rest require manual calculation from raw values, or require some extra processing. Some parameters are applied to the extraction process. Ratio r is calculated with the number of active days, ad, and posts, p, for each student as defined in Eq. 3. We exclude students who have a value of $p < 2$ and a ratio of $r < 0.1$ to refine the dataset to active students.

$$r = p/ad \tag{3}$$

Equation 4 defines the process of calculating a quantifiable measure of ot to investigate how valuable each post is to a student's learning journey. This is achieved through function $f(s, cd)$ performing transformations on the student post body text s and the description text of each course cd.

$$s \rightarrow f(s, cd) \rightarrow ot \tag{4}$$

To construct $f(s, cd)$, we utilise a modification of BERT [35], optimised for NLP transformations on sentences, or Sentence-BERT[2] [36]. Normal BERT maps sentences to a vector space, however, has limitations with common similarity measures. Sentence-BERT overcomes this using a Siamese/triplet network architecture, which improves processing efficiency on big sentences. We use it to convert the post content string (s) and the course description string (cd) sourced from each course's 'about' page into semantic sentence embed values (s_v and cd_v respectively), while also converting s into a semantically structured data item. The distance value is calculated by comparing the course description value with each post and calculate the cosine distance between each using PyTorch's formula [37] outlined in Eq. 5. The process of creating the variables for the proposed model is outlined in Algorithm 1.

$$\text{similarity} = \frac{x_1 \cdot x_2}{\max(\|x_1\|_2 \cdot \|x_2\|_2, \epsilon)} \tag{5}$$

Algorithm 1: Calculation of ot for Each Post s Value

```
Result: ot is calculated for each post in data set
Input: s, cd
for each row in data set do
    Encode s to sentence embed value s_v;
    Encode cd to sentence embed value cd_v;
    Use cosine similarity formula to calculate distance
      between s_v and cd_v;
end
Output: ot
```

The semantic distance score typically ranges between 0 (not semantically similar to the corpus) and 1 (semantically very similar or the same as the corpus). Negative scores in the context of Sentence-BERT are inferred to indicate posts that not only have very little in common with the overall course topic, but also add little-to-no value to the discussion forum. Posts with semantic similarity scores of $ot < 0.02$ were removed, as were short posts (10 words or less) to filter out the 'noise' of introductory or meaningless posts.

To calculate sentiment and stress, the 'BERT-ified' text content, s, is used. Sentiment scores are calculated using the SentiStrength library [38], which has a proven record for providing insight into a user's short informal texts [10], [39]. This treats the 'BERT-ified' post content string s as the input and returns sentiment feature values which we manually add to our dataset. We calculate stress scores based off of the body text of the post made by the student using the library TensiStrength [7]. This is represented in Eq. 6.

$$s \rightarrow f(s) \rightarrow [SEN, STR] \tag{6}$$

[1] AdelaideX: https://www.edx.org/school/adelaidex

[2] Sentence-BERT GitHub repository: https://github.com/UKPLab/sentence-transformers

C. MODEL DEVELOPMENT

Our machine learning objective is to fit a series of numeric feature values, defined as $x = [fv_1, \ldots, fv_n]$, where n is equal the number of features extracted from a data source making up a student-behavior model. In Eq. 7, we run function $f(x)$ and output predictions of the academic performance value associated with that post and student information.

$$x \rightarrow f(x) \rightarrow ap \tag{7}$$

The proposed model, denoted as M, combines engagement, E, semantic, SM, and sentiment, SN, can be formalised as follows in Eq. 8, where the features are combined into a single-layer data model.

$$M = \{[E_1, \ldots, E_n], [SM_1, \ldots, SM_n], [SN_1, \ldots, SN_n]\} \tag{8}$$

We remove elements discovered to have little impact on the importance coefficients when fitting the model – if the results do not change for other features once another is removed, then the element does not impact the rest of the model and is excluded. The baseline models we use to validate our MOOC dataset are: Naíve Bayes (NB), Random Forests (RF) and a Deep Learning Artificial Neural Network (ANN). As stated in Section II, these classifiers are the most commonly used in the educational domain and therefore represent standard classification techniques for experimental testing [6], [31].

For our modelling purposes, we use a Gaussian NB implementation in our technical model. We use Random Forests as a classifier, which implements the Gini index as a means of calculating branch splits. The implementation of RF and NB algorithms are achieved using Sklearn libraries as illustrated in Algorithm 2. We implement an Artificial Neural Network (ANN) for our model using Keras described in Algorithm 3.

Algorithm 2: Creation and Fitting of Machine Learning Model Using NB or RF Algorithms

Result: Fitted machine learning model is created for validation
Input: $M = \langle E, SM, SN \rangle$
1. Create label variable for our model of ap of X;
2. Define set of features for our model from feature sets E, SM, and SN as Y;
3. Create training and testing splits for X and Y $X_{train}, X_{test}, Y_{train}, Y_{test}$;
4. Create model mod and initialise with machine learning algorithm;
5. Fit mod with $X_{train}, X_{test}, Y_{train}, Y_{test}$;
Output: Model mod

Finally, we propose a combined algorithm that takes the best performing predictions from our previous algorithm implementations. The ensemble model uses a combined sample of each algorithm defined above, which allows for more nuanced results and better performance overall across experiments, as illustrated in Algorithm 4.

Algorithm 3: Creation and Fitting of an Artificial Neural Network Model

Result: Artificial Neural Network model is fitted and compiled
Input: $M = \langle E, SM, SN \rangle, k$
1. Create initial model as in Algorithm 2;
2. Add visible and hidden layers to mod;
3. Compile mod with categorical classifier;
4. Initialize compiler with k-fold cross validation;
Output: Model mod

Algorithm 4: Creation and Fitting of a Combined Machine Learning Model

Result: Ensemble model is fitted and compiled
Input: $mod_{NB}, mod_{RF}, mod_{ANN}$
1. Create initial models as in Algorithm 2 for NB and RF algorithms;
2. Create initial ANN model as in Algorithm 3 ;
3. Insert each model in a method that compiles and validates each one in turn;
4. Compare results from each model and keep best performing result;
Output: Model $mod_{combined}$

This combined method follows ensemble machine learning principles of using stacked and voting methods [40]. A stacking method is an aggregate of our models' predictions, taking the best results for features across our models [41]. This allows the strengths of each model to shine and contribute to our prediction service. Comparatively, the voting method of uses a 'majority rule' decision for our predictions not unlike our Random Forests model, but using several models. This generates results using a combined brain of all of our outlined models to make decisions.

V. EXPERIMENT RESULTS

The aim of our experiment is to determine any significant relationship between student posting behavior on discussion forums and final grade, ap. From this, we can determine if student behavior sets measurable and predictable patterns, that can arrive at a particular grade. We use ap as the label for our data model and the remaining features are for prediction. Our testing/training split is 30/70% respectively and we incorporate k-fold cross-validation as described in Algorithm 3, where $k = 5$ to mitigate risk of an unbalanced dataset and investigate performance stability. Experiments were conducted using Python in a Jupyter Notebook environment on a remote university server owned by The University of Adelaide. We utilised the Python libraries: PyTorch, Keras and Tensorflow, Sklearn and Sentence-BERT as described in previous sections. Our performance measuring schemes use industry-standard metrics, accuracy, precision, recall and F1, utilising 5-fold cross validation to generate them.

TABLE 3. Resulting data model statistics.

Cohort	2019 Audit	2019 Verified	2020 Audit	2020 Verified
Post Count	3688	1656	6957	2797
Student Count	1906	363	3479	599
Avg ap	F	D	F	C
Avg ad	6.1	11.86	2.58	10.13
Avg p	5.45	15.81	5.49	12.6
Avg r	1.19	1.65	3.89	4.36
Avg ot	0.3951	0.4088	0.3965	0.3898

TABLE 4. Significance between cohorts.

Measure	Cohorts	p-value
ad	2019 Audit vs 2020 Audit	5.9782E-229
	2019 Verified vs. 2020 Verified	4.5382E-11
p	2019 Audit vs. 2020 Audit	0.374155
	2019 Verified vs. 2020 Verified	1.36343E-14
r	2019 Audit vs. 2020 Audit	1.3699E-167
	2019 Verified vs. 2020 Verified	1.7247E-20
ot	2019 Audit vs. 2020 Audit	0.32029
	2019 Verified vs. 2020 Verified	1.45172E-05

A. RESULTS AND ANALYSIS

From initial experimentation, we observed that whether a student had paid or not played the largest role in representing their behavior. We decided to split the data along this feature (auditing and verified students) to better focus the model on behavioral measures. The resulting model statistics are shown in Table 3, with average scores used each central tendency measure. Average *ap* is calculated by taking the average numeric grade from our data source and fitting it to the grading categories.

We conducted an independent samples t-test to compare significance between the central tendency measures (averages) of each cohort by year. The test was performed under the hypothesis that students engaged as consistently in 2020 as they did in 2019 for (a) auditing students and (b) verified students, equally. The resultant p-values are shown in the far right column of Table 4.

From the p-value results in Table 4, we can observe that there is significance across several of the measures, with p-value < 0.05 for each cohort comparison across *ad* and r. In p and ot, the p-value showed significance among only verified cohorts, while p-value > 0.3 for auditing students. This indicates that engagement behaviors for the 2020 verified cohort was significantly different compared to the 2019 group, therefore, we can reject the null hypothesis that student groups engaged as consistently across all measures in 2020 as they did the previous year. For auditing students, the largest difference was in active days *ad*, meaning there was a clear drop off in simply accessing the course for the 2020 group compared to the 2019 students. A key difference for verified students was in posting ratio r, which indicates that the 2020 group were using the forums far more per *ad* compared to the group in 2019. To understand whether personal emotion might have resulted in these key changes, we then conducted measures of stress and sentiment on the dataset.

TensiStrength is used to extract stress from the dataset. Table 5 shows the values for positive and negative *sen* and *str*, with a slightly greater degree of stress, $-str$, for verified students in both years.

TABLE 5. Resulting data model sentiment statistics.

Cohort	Avg +sen	Avg -sen	Avg $+str$	Avg -str
2019 Audit	2.096	-1.403	2.023	-1.502
2019 Verified	2.062	-1.401	2.056	-1.814
2020 Audit	2.105	-1.394	2.055	-1.549
2020 Verified	2.054	-1.400	2.007	-1.662

TABLE 6. Naive Bayes algorithm performance results.

Cohort	Accuracy	Precision	Recall	F1
2019 Audit	0.6233	0.8475	0.6233	0.7121
2019 Verified	0.5553	0.49	0.5553	0.4544
2020 Audit	0.9822	0.9889	0.9822	0.9853
2020 Verified	0.1954	0.3572	0.1954	0.2107

TABLE 7. Random Forests algorithm performance results.

Cohort	Accuracy	Precision	Recall	F1
2019 Audit	0.9033	0.8627	0.9033	0.86
2019 Verified	0.6036	0.5743	0.6036	0.5535
2020 Audit	0.9942	0.9885	0.9942	0.9913
2020 Verified	0.5125	0.4981	0.5125	0.4651

TABLE 8. ANN deep learning algorithm performance results.

Cohort	Accuracy	Precision	Recall	F1
2019 Audit	0.9174	0.9181	0.9171	0.9175
2019 Verified	0.5406	0.5596	0.3161	0.3786
2020 Audit	0.9932	0.9932	0.9932	0.9932
2020 Verified	0.4308	0.5512	0.1073	0.1644

TABLE 9. Combined voting method of baseline models performance results.

Cohort	Accuracy	Precision	Recall	F1
2019 Audit	0.9098	0.8278	0.9098	0.9528
2019 Verified	0.5121	0.5193	0.9768	0.6781
2020 Audit	0.9935	0.9935	0.9935	0.9903
2020 Verified	0.3887	0.3990	0.6767	0.5020

This model data shows a greater number of total posts in 2020 compared to 2019. Average *ad* reduced for all students in 2020, but r increased significantly. Average overall p is similar between 2019 and 2020, with verified student post numbers reducing slightly in 2020. We validate the *ap* scores using the features through our modelling layer processes, with performance measures for the baseline models outlined in Tables 6, 7 and 8.

Results show high performance measures of > 0.8 for all auditing student cohorts, while verified cohorts have mixed results. Random Forests performed the best out of the baseline models, with all performance metrics reaching approximately 0.5, save for an F1 score for verified students at 0.4651.

In Tables 9 and 10, the results of the voting and stacked method results are compared. Looking purely at verified students, the baseline models generally outperformed the ensemble method in accuracy, however the voting method had higher F1 values with 0.6781 for 2019 verified students and 0.502 for 2020 verified students, which were higher performance than the baseline models and the stacked method. The voting method was able to more accurately return *ap* values for students.

Of the two, the voting method achieved the best performance metrics across the board, out-performing most

TABLE 10. Combined stacked method of baseline models performance results.

Cohort	Accuracy	Precision	Recall	F1
2019 Audit	0.9098	0.9098	0.9098	0.9528
2019 Verified	0.5211	0.5211	0.5211	0.3571
2020 Audit	0.9935	0.9871	0.9935	0.9903
2020 Verified	0.4093	0.3224	0.4093	0.3412

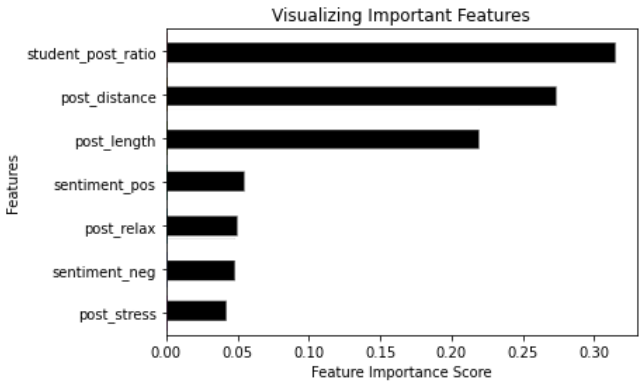


FIGURE 4. Feature importance values for 2019 student cohort. Longer bars indicate a higher importance, with student_post_ratio showing the highest score and post_stress showing the lowest importance score.

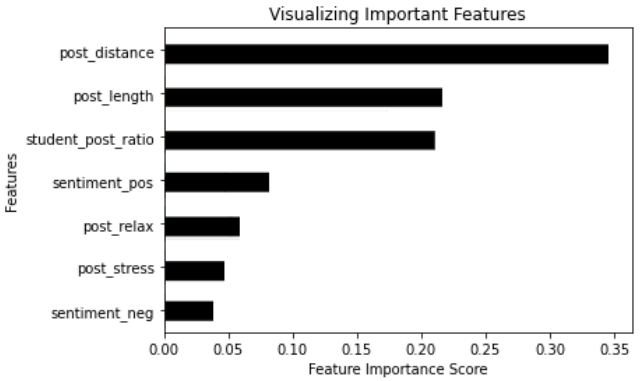


FIGURE 5. Feature importance values for elements in the proposed model for 2020 student. cohort. Longer bars indicate a higher importance, with post_distance showing the highest score and sentiment_neg showing the lowest importance score.

measures found in other algorithms. Our ensemble method appears to largely use RF results. Among the baselines, RF provided best results for accuracy and precision in verified study modes. Naive Bayes appeared to work inconsistently, possibly because it assumes independence among features. We can take this as an indication that features are not independent, and may effect each other. Thus, the RF predictions more closely aligned with actual *ap* values, with accuracy scores > 0.5 for all cohorts.

Finally, we compare the weighting of important features between the two years. Figures 4 and 5 are provided by the Random Forests algorithm, demonstrating which features had the highest importance for *ap* among students. These describe the importance values calculated by our experiments, correlating to $\alpha, \beta, \ldots, \gamma$ in our original research definition formalised in Eq. 2. In 2019, r is weighted heavily in importance compared to other features, while in 2020 it is ranked third. Posting distance, the *ot* value, is ranked in the top two for both years, meaning students posting on topic in the forums had significant impact on their overall outcomes. It is observed that stress and sentiment have lower overall importance for both cohorts, with negative sentiment and negative stress ranked the least important features for both years. From this, there was relatively little change in stress during the pandemic. Engagement features still remained the most important indication of how students would perform in the course.

VI. DISCUSSION

Our overall results demonstrate that of the three categorical features, engagement had the most effect on student academic performance. This is consistent with observations in other similar studies. There were marginal differences in sentiment and stress scores between the two years - in fact, Table 5 demonstrates that 2020 verified students exhibited less negative stress in their posts compared to the cohort in the previous year. Positive sentiment and positive stress scores were very similar across all cohorts, with negative sentiment more or less remaining the same.

As highlighted in Table 3, in 2020 the active days *ad* for all students decreased noticeably, while posting ratios r for all students increased substantially. While students were spending less time on the courses, their time on the forums was up on the previous year as they showed more productive behaviors. However, on topic scores *ot* were more or less the same with previous year averages, so students were not exhibiting more relevant posting behavior in spite of the higher ratios. Results in Table 3 showed that the total number of students in 2020 within our refined dataset was closer to the original number, with 4098 retained against the original number of 4258 after processing. Students from 2019 who 'survived' the pre-processing were much less compared to their original count, with 2269 compared to 4553, meaning over half were lost through our pre-processing. The implications for our ensemble method is that a greater proportion of 2020 students were engaging in a more meaningful way compared to the 2019 cohort. This might suggest that the pandemic may have given verified students more incentive or time to participate in 2020 than the group from 2019.

The differences in physical activity between auditing and verified students was also captured in Table 3. Verified students tended to produce more posts on the forum and have more active days in the courses compared to auditing students, who were particularly down in overall *ad* in 2020 from 2019. This was particularly evident in Table 4, where there was significant differences in the active days of auditing students between the two years, meaning the 2020 group were comparatively less engaged and clearly impacted by something as a whole. This could indicate the difference in priorities during the pandemic for the 2020 cohort, with auditing students unable to engage with online studies due to life circumstances. Verified students were down in *ad* in 2020 compared to the previous year, however, the activity levels were still significant enough to indicate that verified students were far more invested in their outcomes. Table 4

also showed a significant p-value for all engagement measures for verified students, indicating that the 2020 verified students were behaving differently compared to the previous cohort. While verified students generally are expected to participate in a course more by virtue of paying, it was clear that there was far more engagement in the 2020 cohort as a whole. These students may also have been driven to participate more due to unseen factors. As one starting point to further understand this change, future work needs to investigate when active days *ad* occurred for each of these cohorts, for example, whether students accessed the course more during assessment milestones. This would provide insight into what points of a semester verified and auditing students are most likely to be active, and when instructors can alter their discussion forum communication to account for predicted student behavior. This also leads to the first key limitation that the 2019 and 2020 cohorts are not the same collective of students. While we can compare the engagement between these disparate cohorts, it would be useful to compare behaviors across consecutive years using the same student group. This would provide more genuine insight and meaning to our quantitative results to see if there are significant differences in the same group of students as a result of the pandemic.

There was also significant disparity between average grades achieved between verified and auditing cohorts, with all auditing cohorts achieving F compared to C or D as observed in verified cohorts. Auditing students not managing to achieve (or perhaps seeking to achieve) a passing grade is constant across both years, suggesting that the underlying patterns of these students are constant, which is a consistent observation in MOOC cohorts. The unbalanced nature of the dataset may contribute to the performance metrics, with 50% of grades denoted as F, while more granularity is required for passing grades that have a 4-tier spectrum. Additional granularity for failing students may improve this model for greater insight into sub-cohorts, who are presently not well-represented. For example, students who engage with the content but never attempt any assessments are grouped with students who attempt assessment and fail. These two groups are very different and this model may presently be limited by not differentiating between them. This may also indicate that the model remains useful for uneven year or cohort analyses, but this remains a limitation of our approach.

Another key limitation is that although our metrics demonstrate that sentiment and stress do not play a significant role in academic performance, these quantitative results give little insight into student motivation, which is necessary for designing student engagement. Our results indicate that students made use of the forums, with posting on topic demonstrating clear impact on their grades. Posting_distance in Figures 2 and 3, was in the top two across both years. Negative sentiment and stress levels, which were expected to be higher in 2020, showed no change. As stated previously, the interdependence of features may be a useful future investigation. A correlation between posting on-topic actively and reduction in stress may have useful implications for course designers, who can use this as an evidence-base to inform their students that engaging with forums more frequently can have positive effects on their overall well-being. In terms of the ensemble method's potential as a real-time monitoring tool, the use of TensiStrength demonstrates that there is value in detecting stress in conjunction with other categorical features. This can help instructors with not only insight into student interaction, but gather emotional data that will can help understand the overall mood of a large cohort.

VII. CONCLUSION

This work developed an ensemble method for modelling student behavior using features of engagement, semantics and sentiment/stress extracted from a MOOC discussion forum dataset. Our objective was to observe the role of stress in academic performance with a comparison between pre- and during-COVID cohorts as a secondary analysis. The results show that engagement had the most impact on student outcomes, with stress and sentiment rated the least important, even during the pandemic. Addressing the research questions posited in Section I: (1) stress had little impact on academic performance and ranked among the least important features in both years, and (2) stress did not increase during the pandemic, with results indicating its importance decreased compared to 2019. TensiStrength was used for more nuance in understanding stress, which may be useful for MOOC researchers who are improving the potential of real-time monitoring tools.

The work is limited by the selected data range. While we aimed to compare pre- and during-COVID behaviors, one year is perhaps inadequate to formulate an understanding of pre-pandemic behaviors. It was clear that students in 2020 were engaging more actively with the forums compared to the previous year, but whether this was due to the effects of the pandemic remains unknown. Additional analysis should expand the time range selection, to make a comparison between yearly behaviors that would further contextualise the results of 2020. Future work should also utilise more granular analysis to model the behaviors of sets of students within the datasets for refined comparisons. An interesting future endeavour may be to identify a set of students who are represented longitudinally across the course and modelling their student journey, pre- and during-COVID years. As indicated in our Discussion, a more longitudinal, granular analysis that uses the same set of students would provide more contextualised and meaningful insight into the impact of stress and generate a clearer comparison. Nonetheless, our approach to separate sentiment and stress into distinctive features makes a contribution to textual classification studies.

ACKNOWLEDGMENT

The authors would like to thank Prof. Lyn Alderman and the Academic Transformation Portfolio Unit in the University of Southern Queensland for supporting this work. They also extend thanks to Ali Ogilvie and the entire Online Programs Team from The University of Adelaide for approval of data use and their ongoing support of this project. Finally, they would like to thank Mike Thelwall for providing access

to SentiStrength and TensiStrength for the purpose of this research.

AUTHOR CONTRIBUTIONS

Conceptualization, Tao, X.; Shannon-Honson, A.; methodology, Tao, X.; Shannon-Honson, A.; Delaney, P.; validation, Tao, X.; Shannon-Honson, A.; formal analysis, Tao, X.; Shannon-Honson, A.; Delaney, P.; Lin, L.; Dann, C.; Xie, H.; Li, Y.; investigation, Tao, X.; Shannon-Honson, A.; Delaney, P.; Lin, L.; Dann, C.; Xie, H.; Li, Y.; resources, Tao, X.; Shannon-Honson, A.; data curation, Shannon-Honson, A.; writing, Tao, X.; Shannon-Honson, A.; Delaney, P.; Lin, L.; Dann, C.; Xie, H.; Li, Y.; supervision, Tao, X.; project administration, Tao, X. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

[1] A. F. Wise, Y. Cui, and J. Vytasek, "Bringing order to chaos in MOOC discussion forums with content-related thread identification," in *Proc. 6th Int. Conf. Learn. Anal. Knowl.*, 2016, pp. 188–197, doi: 10.1145/2883851.2883916.

[2] O. Almatrafi, A. Johri, and H. Rangwala, "Needle in a haystack: Identifying learner posts that require urgent response in MOOC discussion forums," *Comput. Educ.*, vol. 118, pp. 1–9, Mar. 2018, doi: 10.1016/j.compedu.2017.11.002.

[3] M. Wen, D. Yang, and C. Rosé, "Sentiment analysis in MOOC discussion forums: What does it tell us?" in *Proc. Educ. Data Mining*, Jan. 2014, pp. 130–137.

[4] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, I. Estévez-Ayres, and C. D. Kloos, "Sentiment analysis in MOOCs: A case study," in *Proc. IEEE Global Eng. Educ. Conf. (EDUCON)*, Apr. 2018, pp. 1489–1496, doi: 10.1109/EDUCON.2018.8363409.

[5] K. Lundqvist, T. Liyanagunawardena, and L. Starkey, "Evaluation of student feedback within a MOOC using sentiment analysis and target groups," *Int. Rev. Res. Open Distrib. Learn.*, vol. 21, no. 3, pp. 140–156, May 2020, doi: 10.19173/irrodl.v21i3.4783.

[6] Z. Kastrati, F. Dalipi, A. S. Imran, K. Pireva Nuci, and M. A. Wani, "Sentiment analysis of students' feedback with NLP and deep learning: A systematic mapping study," *Appl. Sci.*, vol. 11, no. 9, p. 3986, Apr. 2021, doi: 10.3390/app11093986.

[7] M. Thelwall, "TensiStrength: Stress and relaxation magnitude detection for social media texts," *Inf. Process. Manage.*, vol. 53, no. 1, pp. 106–121, Jan. 2017, doi: 10.1016/j.ipm.2016.06.009.

[8] G. Ilieva, T. Yankova, S. Klisarova-Belcheva, and S. Ivanova, "Effects of COVID-19 pandemic on university students' learning," *Information*, vol. 12, no. 4, p. 163, Apr. 2021, doi: 10.3390/info12040163.

[9] B. K. Dhar, F. K. Ayittey, and S. M. Sarkar, "Impact of COVID-19 on psychology among the university students," *Global Challenges*, vol. 4, no. 11, Nov. 2020, Art. no. 2000038, doi: 10.1002/gch2.202000038.

[10] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 1, pp. 163–173, Jan. 2012, doi: 10.1002/asi.21662.

[11] C. Karyotis, F. Doctor, R. Iqbal, A. James, and V. Chang, "A fuzzy computational model of emotion for cloud based sentiment analysis," *Inf. Sci.*, vols. 433–434, pp. 448–463, Apr. 2018, doi: 10.1016/j.ins.2017.02.004.

[12] A. Kumar and G. Garg, "Sentiment analysis of multimodal Twitter data," *Multimedia Tools Appl.*, vol. 78, no. 17, pp. 24103–24119, Sep. 2019, doi: 10.1007/s11042-019-7390-1.

[13] X. Zhou, X. Tao, M. M. Rahman, and J. Zhang, "Coupling topic modelling in opinion mining for social media analysis," in *Proc. Int. Conf. Web Intell.*, New York, NY, USA, Aug. 2017, pp. 533–540, doi: 10.1145/3106426.3106459.

[14] A. Giachanou and F. Crestani, "Tracking sentiment by time series analysis," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2016, pp. 1037–1040, doi: 10.1145/2911451.2914702.

[15] X. Zhou, X. Tao, J. Yong, and Z. Yang, "Sentiment analysis on tweets for social events," in *Proc. IEEE 17th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, Jun. 2013, pp. 557–562, doi: 10.1109/CSCWD.2013.6581022.

[16] S. Sidana, S. Amer-Yahia, M. Clausel, M. Rebai, S. T. Mai, and M.-R. Amini, "Health monitoring on social media over time," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 8, pp. 1467–1480, Aug. 2018, doi: 10.1109/TKDE.2018.2795606.

[17] X. Tao, X. Zhou, J. Zhang, and J. Yong, *Sentiment Analysis for Depression Detection on Social Networks* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10086. Gold Coast, QLD, Australia: Springer, 2016, pp. 807–810, doi: 10.1007/978-3-319-49586-6_59.

[18] M. A. Alsheri, L. M. Alrajhi, A. Alamri, and A. I. Cristea. (Sep. 10, 2021). *MOOCSent: A Sentiment Predictor for Massive Open Online Courses*. Accessed: Oct. 30, 2020. [Online]. Available: https://aisel.aisnet.org/isd2014/proceedings2021/methodologies/13/

[19] C.-W. Shen and C.-J. Kuo, "Learning in massive open online courses: Evidence from social media mining," *Comput. Hum. Behav.*, vol. 51, pp. 568–577, Oct. 2015, doi: 10.1016/j.chb.2015.02.066.

[20] X. Xie, S. Ge, F. Hu, M. Xie, and N. Jiang, "An improved algorithm for sentiment analysis based on maximum entropy," *Soft Comput.*, vol. 23, no. 2, pp. 599–611, Jan. 2019, doi: 10.1007/s00500-017-2904-0.

[21] R. Gaspar, C. Pedro, P. Panagiotopoulos, and B. Seibt, "Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events," *Comput. Hum. Behav.*, vol. 56, pp. 179–191, Mar. 2016, doi: 10.1016/j.chb.2015.11.040.

[22] Z. Liu, C. Yang, S. Rüdian, S. Liu, L. Zhao, and T. Wang, "Temporal emotion-aspect modeling for discovering what students are concerned about in online course forums," *Interact. Learn. Environments*, vol. 27, nos. 5–6, pp. 598–627, Aug. 2019, doi: 10.1080/10494820.2019.1610449.

[23] R. Gopalakrishna Pillai, M. Thelwall, and C. Orasan, "Detection of stress and relaxation magnitudes for tweets," in *Proc. Companion Web Conf. Web Conf.*, 2018, pp. 1677–1684, doi: 10.1145/3184558.3191627.

[24] S. C. Guntuku, A. Buffone, K. Jaidka, J. C. Eichstaedt, and L. H. Ungar, "Understanding and measuring psychological stress using social media," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 13, no. 1. Palo Alto, CA, USA: AAAI Press, 2019, pp. 214–225.

[25] R. Gopalakrishna Pillai, M. Thelwall, and C. Orasan, "Trouble on the road: Finding reasons for commuter stress from tweets," in *Proc. Workshop Intell. Interact. Syst. Lang. Gener.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 20–25, doi: 10.18653/v1/w18-6705.

[26] X. Chen, M. Sykora, T. Jackson, S. Elayan, and F. Munir, "Tweeting your mental health: An exploration of different classifiers and features with emotional signals in identifying mental health conditions," in *Proc. 51st Hawaii Int. Conf. Syst. Sci.*, 2018, p. 3321, doi: 10.24251/HICSS.2018.421.

[27] K. Sailunaz and R. Alhajj, "Emotion and sentiment analysis from Twitter text," *J. Comput. Sci.*, vol. 36, Sep. 2019, Art. no. 101003, doi: 10.11575/PRISM/32714.

[28] X. Wei, H. Lin, L. Yang, and Y. Yu. "A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification," *Information*, vol. 8, no. 3, p. 92, Jul. 2017, doi: 10.3390/info8030092.

[29] N. Öztürk and S. Ayvaz, "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis," *Telematics Informat.*, vol. 35, no. 1, pp. 136–147, Apr. 2018, doi: 10.1016/j.tele.2017.10.006.

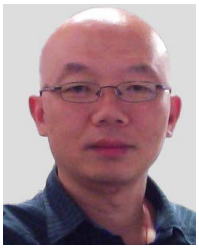
[30] T. Sinha and J. Cassell, "Connecting the dots: Predicting student grade sequences from bursty MOOC interactions over time," in *Proc. 2nd ACM Conf. Learn. Scale*, Mar. 2015, pp. 249–252, doi: 10.1145/2724660.2728669.

[31] K. Mite-Baidal, C. Delgado-Vera, E. Solís-Avilés, A. H. Espinoza, J. Ortiz-Zambrano, and E. Varela-Tapia, "Sentiment analysis in education domain: A systematic literature review," in *Proc. Int. Conf. Technol. Innov.* Cham, Switzerland: Springer, 2018, pp. 285–297, doi: 10.1007/978-3-030-00940-3_21.

[32] The University of Adelaide. (2020). *The University of Adelaide Marks and Grading Scheme*. Accessed: Oct. 30, 2020. [Online]. Available: https://www.adelaide.edu.au/policies/700/?dsn=policy.document;field=data;id=1044;m=view

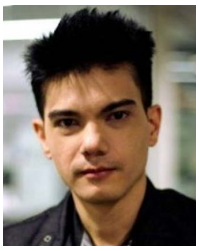
[33] M. Birjali, A. Beni-Hssane, and M. Erritali, "Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks," *Proc. Comput. Sci.*, vol. 113, pp. 65–72, Jan. 2017, doi: 10.1016/j.procs.2017.08.290.

[34] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer, "Forecasting the onset and course of mental illness with Twitter data," *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, Dec. 2017, doi: 10.1038/s41598-017-12961-9.
[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[36] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
[37] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *Proc. NIPS-Workshop*, 2017, pp. 1–4. [Online]. Available: https://openreview.net/forum?id=BJJsrmfCZ
[38] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Assoc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, 2010, doi: 10.1002/asi.21416.
[39] M. Thelwall and K. Buckley, "Topic-based sentiment analysis for the social web: The role of mood and issue-related words," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, no. 8, pp. 1608–1617, 2013, doi: 10.1002/asi.22872.
[40] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Int. Workshop Multiple Classifier Syst.* Cham, Switzerland: Springer, 2000, pp. 1–15, doi: 10.1007/3-540-45014-9_1.
[41] M. Graczyk, T. Lasota, B. Trawiński, and K. Trawiński, "Comparison of bagging, boosting and stacking ensembles applied to real estate appraisal," in *Proc. Asian Conf. Intell. Inf. Database Syst.* Cham, Switzerland: Springer, 2010, pp. 340–350, doi: 10.1007/978-3-642-12101-2_35.



XIAOHUI TAO (Senior Member, IEEE) received the Ph.D. degree from the Queensland University of Technology, Brisbane. He is currently an Active Researcher in AI and an Associate Professor (computing) with the School of Mathematics, Physics and Computing (SoMPC), University of Southern Queensland (USQ), Australia. His research interests include data analytics, machine learning, knowledge engineering, information retrieval, and health informatics. His research outcomes have been published on many top-tier journals (e.g., TKDE, IPM, KBS, ESWA, and PRL) and conferences (e.g., IJCAI, ICDE, CIKM, PAKDD, and WISE). He received the ARC DP Grant (Ref. DP220101360) (2022–2024), an Australian Endeavour Research Fellow from 2015 to 2016, and was awarded with the "Research Performance Award" and the "Discipline Research Performance Improvement Award" by SoMPC, USQ, among many others. He is an elected Senior Member of ACM. He has been active in professional services. He has served as the PC Chair in WI '17 and '18, WI-IAT '21, and BESC '18 and '21 and an Editor or a Guest Editor for many journals, including INFFUS and WWWJ.

AARON SHANNON-HONSON is currently pursuing the master's degree in information technology from the University of Southern Queensland, Australia. He is also working as a member of the Online Programs Team, The University of Adelaide, Australia. Before then, he was a Learning Resource Developer who has worked on the Shakespeare Matters MOOC and the Micromasters Project in The University of Adelaide.



PATRICK DELANEY received the bachelor's degree in information technology and the Ph.D. degree from the Queensland University of Technology, Australia. He currently works with the School of Mathematics, Physics and Computing, University of Southern Queensland. His research interests include data analytics in education, using mixed methodologies and student success, engagement, and retention.

LIN LI (Member, IEEE) received the Ph.D. degree from The University of Tokyo, Japan, in 2009. She is currently a Professor with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology. She has more than 100 publications, including those on AAAI, ICDM, ICMR, CIKM, TOIS, and TOIT have received more than 1300 citations on Google Scholar. Her research interests include information retrieval, recommender systems, data mining, multimedia computing, and natural language processing.

CHRISTOPHER DANN is an inclusive goal orientated leader whose purpose is to make a positive impact on the educational experiences of learners "glocally." He is currently a Senior Lecturer with the School of Teacher Education, University of Southern Queensland, Curriculum and Pedagogy (Technologies). His current research is exploring the possible impact of machine learning and artificial intelligence on the teaching and learning process from the perspective of teachers and students across educational context.

YAN LI is currently a Professor in computer science with the School of Mathematics, Physics and Computing, University of Southern Queensland, Australia. Her research interests include artificial intelligence, big data analytics, signal and image processing, biomedical engineering, and computer networking technologies and security.

HAORAN XIE (Senior Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong. He is currently an Associate Professor with the Department of Computing and Decision Sciences, Lingnan University, Hong Kong. He has published 258 research publications, including 129 journal articles. Among all 129 journal articles, there are 104 SCI/SSCI indexed and 13 SCOPUS indexed. His research interests include artificial intelligence, big data, and educational technology. He has obtained 14 research awards and five best/excellent paper awards from international conferences, including WI 2020, ICBL 2020, DASFAA 2017, ICBL 2016, and SECOP 2015. He has ranked as the world's top 20 researchers in artificial intelligence in education in Google Scholar. His proposed LSGAN, published in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and ICCV, with more than 1000 citations in two years, has been included in the computer vision course at Stanford University and implemented by Google. He has successfully obtained more than 50 research grants; the total amount of these grants is more than HK$27 million. He is the Senior Member of ACM.