Exploring the Limits of Lexicon-based Natural Language Processing Techniques for Measuring Engagement and Predicting MOOC's Certification

Esther Félix, Nicolas Hernandez, Issam Rebaï

Exploring the limits of lexicon-based Natural Language Processing techniques for measuring engagement and predicting MOOC's certification

Esther Félix[1,2], Nicolas Hernandez[1] and Issam Rebaï[2]
[1]*LS2N, Université de Nantes, France*
[2]*Lab-STICC, Institut Mines-Télécom Atlantique, France*
efelix@ensc.fr, nicolas.hernandez@univ-nantes.fr, issam.rebai@imt-atlantique.fr

Keywords: MOOC, Forum Discussions, Engagement, Text Mining, Lexicon-based approach, Learning outcome prediction

Abstract: We address the problem of assessing the contributions of lexicon-based Natural Language Processing (NLP) techniques to measure learner affective and cognitive engagement and thus predict certification in French-speaking MOOCs. Interest in these approaches comes from the fact they are explainable. Our investigation protocol consists of applying machine learning techniques to determine the relationships between lexicon-based engagement indicators and learning outcomes. The lexicon-based approach is compared with trace log features, and we distinguish between specialised linguistically-based approaches with dedicated lexicon resources and more general but deeper text representations. Language quality and its impact on the task are discussed. We investigate this issue in MOOCs imposing or not the use of the forum in their learning activities.

1 INTRODUCTION

Remote learning, whatever its type (MOOC, online training, etc.) delivered by a team of trainers, suffers from the problem of social distancing. On the one hand, learners find themselves isolated from each other. On the other hand, trainers cannot see the learners,their behaviour or their non-verbal communication to probe the learning situation. To overcome these drawbacks, learning platforms have integrated communication tools such as forums and have developed mechanisms for collecting traces in order to feed dashboards for monitoring learners and the training. These dashboards have become an indispensable tool for trainers, training institutions and authors of educational resources. Several research studies have analysed the collected traces, developed machine learning algorithms to feed the dashboards with statistical data or make predictions of success or dropout (Moreno-Marcos et al., 2020). While several studies exist on the analysis of traces collected by the platforms (Ferri, 2019), very few have focused on the linguistic analysis of the content of forums and even fewer studies have combined the analysis of traces with the linguistic analysis of the content of the communication tools (Joksimović et al., 2018; Fincham et al., 2019). By accident, the bulk of the published work has only focused on English-speaking MOOCs.

In this paper, we address the problem of assessing the contributions of lexicon-based Natural Language Processing (NLP) techniques for measuring learner engagement (emotional and cognitive) and for predicting certification in Francophone MOOCs. Our interest for this study stems from the fact that the behaviours of tools that incorporate such techniques are interpretable by humans (Danilevsky et al., 2020). We investigate this issue in two different MOOCs. One of them has the particularity to impose the use of the forum in some of its learning activities. Based on the literature, we explore various linguistically-based approaches to measure individual engagement indicators in MOOCs' forums and to evaluate their relations with learning outcomes. We first review the literature on measuring engagement in the context of MOOCs. We then present our data, the engagement indicators implemented and the NLP text representation models we developed to conduct our investigation. Eventually, we report our experiments on prediction graduation given various combinations of trace logs and linguistic information. We conclude with a discussion on the benefits and limitations of NLP in predicting success and measuring learner engagement.

2 STATE OF THE ART

The notion of engagement is at the crossroads of different fields such as psychology, education and human learning. By validating and extending the engagement model of Reschly and Christenson (2012) to MOOCs, Joksimović et al. (2018) and Fincham et al. (2019) offer a structured and relatively complete modeling of the notions related to the student engagement (i.e. academic, behavioral, cognitive, and affective engagements), and the types of metrics that it is possible to associate. All the various dimensions of engagement are possible to analyze from MOOCs data, but the use of NLP techniques only seems relevant for the cognitive and emotional dimensions, of which indicators are accessible *via* the discussion forums, even if the cognitive dimension is possibly also accessible from other types of written productions of learners (Joksimović et al., 2018; Fincham et al., 2019).

To capture the *affective engagement*, Wen et al. (2014b) use lexicons both to recognize specific MOOC subjects and the sentiment polarity associated to these subjects. They show a correlation between the collective opinion and the dropout phenomena, but they do not observe any influence of the sentiment expressed by a student on his desertion. Like Ramesh et al. (2013), they note that, although positive words can be considered as engagement markers, negative words are not a sign of lack of engagement. Tucker et al. (2014) also use a weighted sentiment lexicon and show a strong negative correlation between sentiments expressed in the discussion forum and the students' average grade, but a positive (but weak) correlation between sentiments and assignment scores. Yang et al. (2015) show that the combination of trace log features (such as click counts) and lexical features can train a logistic regression classifier to determine with success the level of confusion of a learner in his posts. To do so, the authors used the LIWC (Linguistic Inquiry and Word Count) dictionary and its semantic categories of words (Tausczik and Pennebaker, 2010). Their study shows that the dropout rate and the level of confusion are related.
For measuring the *cognitive engagement*, Wen et al. (2014a) propose to use linguistic indicators based on the recognition of specific domain independent keywords in the forum posts such as the apply words (which convey the practice of the lessons), the need words (which can mean the needs and the motivation of the learner), first person words (which can indicate the direct commitment of the learners in their discourse), LIWC-cognitive words (reporting cognitive mechanisms). In the context of modeling Math Identity and Math Success, Crossley et al. (2018) study the complexity and the abstraction levels of the learners' written productions by analysing their lexical and syntactic sophistication, the text cohesion and the expression of sentiments and cognition processes. The study of Atapattu et al. (2019) focuses on the observation of the learner's cognitive engagement in terms of active and constructive behaviours in MOOCs (does the learner handle the course material? Does he create new content?). The authors use word embedding techniques with Doc2vec (Le and Mikolov, 2014) to model the text course material and the learners text productions, and eventually apply vector similarity measures to compare the various text contents. Fincham et al. (2019) integrate several metrics to capture the engagement on several dimensions. In addition to trace logs features, they characterize each learner thanks to median measures in their posts in terms of sentiment polarity, specific emotions (thanks to the IBM Tone Analyzer), and lexical, syntactic and text sophistication and cohesion (thanks to the Coh-Metrix tool). The study shows that correlations exist between these features accounting for the engagement dimensions proposed by Joksimović et al. (2018) model.

So far, all the academic literature report studies which only use data in English and so specialized tools for English processing. This underlines the importance of, first, reproducing the measurements made from English MOOCs on French content. Dominant techniques have based their approach on surface analysis using linguistic resources such as specialized lexicons. The most recent one tends to investigate the interest of using more deep analysis.

3 Investigation protocol

In this section, we present our data, the engagement indicators we implemented, and the measure instruments we handle to observe the influence of these indicators on the learning outcomes.

3.1 Data

Our study takes advantage of gaining access to MOOCs in two different domains: 2 editions of "Digital Fabrication[1]" (*df*) and 3 editions of "French as a Foreign Language Learning[2]" (*ffl*). The MOOCs are held on the FUN online learning platform (a French

[1] https://www.fun-mooc.fr/fr/cours/sinitier-a-la-fabrication-numerique/

[2] https://lms.fun-mooc.fr/courses/course-v1:univnantes+31001+session03/

public interest grouping) with resources (including participants' contents) released under Creative Commons. The *df* editions occurred for 4 weeks. 7.5k persons enrolled the MOOCs, 300 participants posted at least a message (4%) and 1k were graduated (13%). The *ffl* editions occurred for 6 weeks. 11.5k persons enrolled the MOOCs, 1.1k participants posted at least a message (9.6%) and 340k were graduated (3%). For the *ffl* MOOCs, the forum discussions were a compulsory step where learning activities took place. The data follows the format of the EDX tracking logs[3]. We parsed these logs to obtain all those related to the publication of a message in the forum, and kept the content of the messages and the date of their publication in the forum to build our datasets. In addition to the trace logs, we also hold intermediate grades, final grades, and MOOC validations.

3.2 Engagement indicators

The indicators were selected in order to individually depict the profile and the behaviour of each learner. To obtain linguistic-based information for each learner, we concatenated the messages he/she posts into a unique text unit, we call the *user contribution*.

3.2.1 Trace log event-based indicators

Our selection of event-based indicators fits the works of Whitehill et al. (2015); Crossley et al. (2016). The following indicators were implemented: "*Forum interactions counter*": the count of interactions a learner has with the MOOC's forum. Were considered as an interaction: a forum search, a message post, a vote for a message... "*Navigation-click counter*": the count of clicks performed by a learner to switch to another MOOC web page; "*Videos played counter*": the count of videos played by the learner; "*Graded problems counter*": the count of assignments submitted by the learner.

3.2.2 Linguistically-based indicators related to the affective engagement

In contrast with the work from Fincham et al. (2019) which benefits from the availability of the IBM Tone Analyzer for processing student emotions in English, there is no such resource yet for processing emotions in French. Based on a scored-lexicon with polarity and subjectivity values, *TextBlob* offers "*Sentiment polarity*" and "*Text subjectivity*" measures by "averaging" the scores of the occurring lexicon entries in a given text. *TextBlob* provides a French language support[4] (with 5,116 inflected forms) .

3.2.3 Linguistically-based indicators related to the cognitive engagement

We based our approach on Wen et al. (2014a). One of our contribution was to develop dedicated resources for processing French. "*Apply words*": We literally translated the lexicon used by Wen et al. (2014a) thanks to the online *Larousse* dictionary[5] by systematically including all the proposed translations (23 lemmas); "*Need words*": We performed the same protocol as we did with the *apply words* (24 lemmas); "*First person words*": We simply listed the personal pronouns, the possessive pronouns and the possessive adjectives used in French (15 lemmas); "*LIWC-cognitive words*": We used the French LIWC version (Piolat et al., 2011) through the Python module *liwc*[6] (749 lemmas).

3.2.4 Non-specialized linguistically-based indicators

The indicators presented in the two previous sections have been defined by human experts. However, it is possible that engagement marks may be present in forum posts that these metrics do not capture. We therefore decided to experiment more general methods allowing to analyze the whole text, to possibly detect other indicators related to the success that would be present there. We studied three word and text representations: (1) The *Bag-Of-Word (BOW) representation with a TF.IDF word scoring* represents a text by a vector of the words it contains. *TF.IDF* stands for *term frequency*inverse document frequency*. The scoring measures how significant the words are given their frequencies in the text and their frequencies in a whole corpus. The vector dimension size corresponds to the size of the corpus vocabulary. Such approach is quite basic and provides sparse text representations. (2) The *FastText word embedding representation* (Bojanowski et al., 2016)[7] is an extension of the Word2vec skip-gram model (Mikolov et al., 2013) which is a two-layer neural networks (one single hidden layer) where the distributed representation of the input word is used to predict the context

[3] https://edx.readthedocs.io/projects/devdata/en/stable/internal_data_formats/tracking_logs.html

[4] https://github.com/sloria/textblob-fr

[5] https://www.larousse.fr/dictionnaires/bilingues

[6] https://pypi.org/project/liwc

[7] Developed by Facebook https://fasttext.cc

(the surrounding words). It is self-supervised learning i.e. it does not require any labeling effort for building the training data. Weights of the hidden layer are learned by observing the words in their context in a corpus. Eventually they correspond to the "word vectors" the model learns. While Word2vec takes words as input, FastText processes substrings of words (character n-grams). This ability allows it to build vectors even for misspelled words or concatenation of words. Compared to *TF.IDF* which produces a score per word, FastText produces a finer representation by providing one vector per word. A sentence/document vector is obtained by averaging the word/ngram embeddings. (3) *The BERT Language representation Model*[8] (Devlin et al., 2018) aims at learning the probability distribution of words in a language. BERT stands for Bidirectional Encoder Representations from Transformers. Such approaches do not produce static word embeddings (like word2vec approaches) but produce contextualized word embeddings which are a finer representation of text content. For our experiments, we use the Multilingual Cased model. Roughtly speaking, BOW with *TF.IDF* can be considered as the old-fashioned approach to model texts in Natural Language Processing, Word Embeddings are the dominant approach in the last decade while Language models are at the cutting edge.

3.3 Measuring instruments

We applied machine learning techniques to determine the possible relationships between the indicators and the learning outcomes as well as the weights of each feature involved in the engagement prediction (i.e. the learning outcomes). We defined the prediction of the success in graduation of the MOOC participants as a binary classification problem. To determine the influence of the event-based and specialized linguistic-based indicators in the graduation prediction we used a logistic regression algorithm which offers the advantage of requiring very little runtime to operate as well the ability to estimate the individual influence of each feature. We also use the same model with the general linguistic-based indicators built with a bow and *TF.IDF* scoring. Concerning the FastText word embeddings, the original library offers an implementation of the architecture with an additional layer which uses a multinominal logistic regression for handling classification tasks (Joulin et al., 2016). The sentence/document vector corresponds so to the features. In a similar way, pre-trained BERT models can be fine-tuned with just one additional output layer to create models for a wide range of tasks, such as classification task. In the next section, we report experiments combining and mixing our data. Our objective is to compare the prediction performance of the models built with various feature configurations in input: (1) Event-based indicators as features, (2) combination of event-based and linguistic indicators and (3) linguistic indicators on their own. The general procedure for experimenting was to train a model on 80% of the data (randomly selected), then to evaluate the performance of the machine learning models on the 20% data remaining. To do so, we used the following instruments and metrics: Confusion Matrix, Accuracy, Precision, Recall, F1 Score. For the BERT and FastText models, we also used a validation set (the test set was split on purpose).

4 Corpus analysis

In order to be able to discuss the results of the experiments reported in Section 5, we conducted some brief studies to assess the French language quality in our corpus as well as to measure the presence of our lexicon-based indicators in the corpus.

4.1 French language quality assessment

Since the *ffl* MOOCs were written by non native French speaker, it is right to assess the quality of the users contributions. To do so, we observed three kinds of measures: the coverage of detected languages, the pseudo-perplexity (PPPL) metric and the coverage of a French lexicon.

4.1.1 Language detection

Thanks to the Compact Language Detector 2 (CLD2)[9], we detected and computed the proportion of each identified language in each MOOC. The CLD2 detection mainly relies on the probability to observe 4-characters-grams for each known language. For both types of corpus, the distribution is homogeneous over all the editions. For *ffl*, French represents about 96% percent of texts, arround 2% percents are unknown while the remaining 2% percents covers up to 5 distinct other languages (English for 0.6%). For *df*, about 82% of the posts are written in French, 12 % are in English and the 8% remaining count as unknown (mainly programming language).

[8] Developed by Google `https://github.com/google-research/bert`, we used the *ktrain* framework to interface BERT `https://github.com/amaiya/ktrain`

[9] `https://github.com/aboSamoor/pycld2`

4.1.2 Language Model Pseudo-Perplexity

Pretrained language models are commonly used in NLP tasks (e.g. machine translation, speech recognition) to estimate the probability of a word sequence and Perplexity (PPL) is a traditional intrinsic metric to evaluate how well a model can predict the word sequence of an unseen text (Martinc et al., 2021). The lower the PPL score is, the better the language model predicts the words in a text. We benefited from the freely available neural language models pretrained for French and we used a PPL version adapted to evaluate neural language model namely the pseudo-perplexity (PPPL) proposed by Salazar et al. (2020). As a language model we used an instance of the popular Generative Pre-trained Transformer 2 (GPT2) model[10].

For each MOOC, we computed the PPPL of the 100 first tokens of each user contributions, and report the mean, median and standard deviation. As a reference, we also computed the PPL of 1000 sentences randomly selected from French texts of the Gutenberg project (sentence length greater or equal than 5 whitespace-separated tokens). The GPT2 model was partially trained with the Gutenberg project[11].

Table 1: GPT2 pseudo-perplexity. Contributions count, PPPL mean, median and standard deviation.

corpus	contrib.	mean	median	std
ffl1	781	114.93	57.61	395.23
ffl2	1176	108.43	56	360.04
ffl3	788	648.67	52.6	11694.7
df1	230	221.12	77.94	770.64
df2	248	507.95	71.15	6004.89
df1-fr	208	113.26	71.49	161.83
df2-fr	228	101.8	67.34	94.22
gutenberg	1000	67.97	29.63	223.78

In Table 1, *df1-fr* and *df2-fr* correspond respectively to a version of *df1* and *df2* with the detected French text parts. Mean is difficult to interpret because it erases the differences but we note that medians are homogeneous for both types of MOOCs (*ffl* and *df*) and the *ffl* PPPLs are lower than the `df` ones, which means that the GPT2 model predicts more easily the words sequence of *ffl* than *df*. We also note that even if the Gutenberg PPPL median is almost twice lower, the step is not than large.

[10] https://huggingface.co/asi/gpt-fr-cased-small

[11] https://www.gutenberg.org

4.1.3 French lexicon coverage

Lastly, we checked the proportion of MOOC's vocabulary belonging to French. We merge the glaff[12] (fr.wiktionnaire.org) and the lefff[13] lexicons as a base to cover all the derivational and inflectional forms existing in French (1,177,561 entries). Lowercase and tokenization with spacy[14] were performed as preprocessing.

Table 2: French lexicon coverage. Percentage of MOOC vocabulary out of the French lexicon (OOV entry) and percentage of word occurrences in full text which are out of the French lexicon (OOV occ.).

corpus	% OOV occ.	% OOV entry
ffl1	9.46	31.11
ffl2	13.28	40.11
ffl3	11.66	36.06
df1	23.86	26.92
df2	21.76	26.47

In Table2, for *ffl* MOOCs, we observed that about 35 % of vocabulary are out of our French Lexicon and the OOV words occur 1 word over ten in full text. For *df*, we observed that about 27 % of vocabulary are out of our French Lexicon and these OOV words occur almost once over 4 words.

The *ffl* OOV words are foreign words, misspelled words, first names and wrongly tokenized words. The *df* OOV words are mainly language programming terms or subwords.

The three studies in this section tend to show that the language quality of the *ffl* MOOCs is good or at least not less than the *df* MOOCs written by native French speakers.

The question arises as to whether the semantic lexicons used to build the linguistic features are represented in our data.

4.2 Coverage of the linguistic features in the corpus

In Table 3, we observe that all semantic classes are present. There is a difference between *ffl* and *df* figures (*df* figures are lower) but the trends are similar. Apply words are used by 45% of the posting users in *ffl* and about 25% in *df*. Need words are used by 50% of the posting users for all MOOCs. First words are used by 90% of the posting users for all MOOCs. LIWC-cognition words are used by about 96% of the posting users for all MOOCs. TextBlob sentiment

[12] http://redac.univ-tlse2.fr/lexicons/glaff_en.html

[13] http://pauillac.inria.fr/~sagot/

[14] https://spacy.io/

Table 3: Coverage of the linguistic features in the corpus the number of users, the number of posting users (at least one post), the absolute (abs.) count of occurrences of a given linguistic feature (apply, need, first-person, LIWC-cognition and TextBlob sentiment), the relative (rel.) count of posting users with at least one occurrence of the given linguistic features.

			apply words		need words		first-person		LIWC-cognition		tb-sentiment	
corpus	users	posting	abs.	rel. (%)	abs.	rel. (%)	abs.	rel. (%)	abs.	rel. (%)	abs.	rel. (%)
ffl1	3595	781	885	357 (46)	1164	407 (52)	10615	727 (93)	34933	763 (98)	19294	752 (96)
ffl2	4859	1176	891	433 (37)	1027	484 (41)	11426	1129 (96)	23257	1105 (94)	12779	1030 (88)
ffl3	4109	788	1029	408 (52)	1359	449 (57)	10526	751 (95)	29916	764 (97)	17240	761 (97)
df1	4051	230	110	57 (25)	147	80 (35)	1350	191 (83)	5497	218 (95)	2491	207 (90)
df2	2569	248	129	65 (26)	193	97 (39)	1841	212 (85)	6108	239 (96)	3008	228 (92)

words are used by about 93% of the posting users for all MOOCs. Absolute counts of first person, LIWC-cognition and TextBlob-sentiment occurrences show a large number of occurrences of these classes (11 times the number of posting users for first person, 28 times for LIWC-cognitive and 17 times for TextBlob-sentiment).

At this stage, the question arises of the selective capacity of apply and needs words because of the resources scarcity, and similarly for first-person and LIWC-cognition words due to their over-representation in the data. The presence of sentiment words means that TextBlob should have material to measure subjectivity and sentiment polarity.

5 Experiments

The purpose of the experiments was to observe a relationship between linguistic indicators and learner outcomes. First, we performed simple correlation calculations between linguistically-based indicators and learners' grades (with Pearson, Spearman and distance correlations coefficients). We did not find conclusive correlation results. We then used logistic regression models with event-based indicators and linguistically-based indicators as features to predict success or failure to the MOOC, in order to determine if a link exists between those features and academic success.

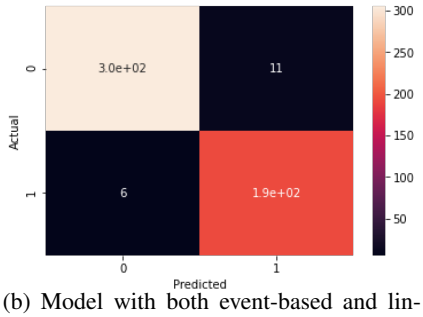
5.1 First experiment: Contribution of the linguistically-based indicators to the event-based indicators

Our goal was to compare the prediction results obtained with three distinct models: A model taking event-based indicators as input features, a model combining event-based and linguistically-based indicators, and a model taking only linguistically-based indicators. The event-based features used are those defined in section 3.2.1. For this experiment, the linguistic features chosen were the linguistically-based indicators related to the cognitive and affect engagement, applied, for each learner, to the concatenation of their messages.

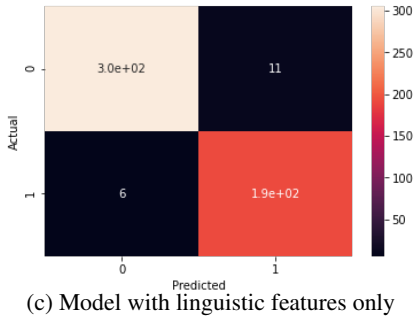
We experiment on the five MOOC editions (*df1, df2, ffl1, ffl2* and *ffl3*) as datasets. Since the results of the experiments on each *df* dataset were similar between eachother, respectively on each *ffl* dataset, we only show the results for one of each dataset. Fig-

Figure 1: Confusion matrices corresponding to the predictions of logistic regression models for *df2* data

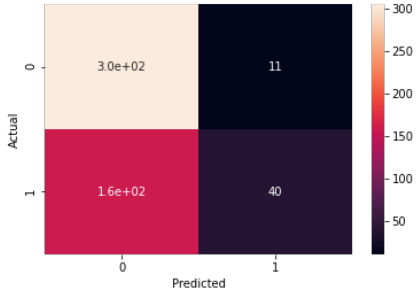
(a) Model with event-based features only



(b) Model with both event-based and linguistic features



(c) Model with linguistic features only



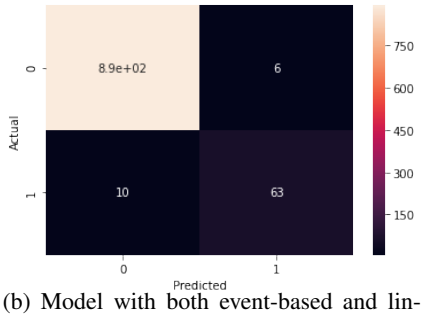
ure 1 gives the prediction results for *df2* data and Figure 2, from *ffl2* data, in each of the previously mentioned configurations. They show the raw number of students who were correctly and incorrectly classified as successful or unsuccessful, based on the indicators given as input to the classifier. Tables 4 and 5 give the associated accuracy, precision, recall and F1-score.

We observe that the results of the models trained with event-based features obtain excellent prediction results. On the other hand, the addition of linguistically-based features seems to add nothing to the model since the results obtained are identical. We were able to check that the students predicted by the models as having failed were the same for the models with and without linguistically-based features. Concerning the results of the models trained only with linguistic features, they seem to confirm that the chosen linguistic features do not help the prediction, at least in this experiment. Indeed, in this configuration, for both *df2* and *ffl2*, the model classifies almost all students in the "failure" category. This is probably due to the imbalance in proportions between failing and successful students: since there are more failing students, putting them all in this category allows the model to obtain a good accuracy, at the cost of precision and recall.

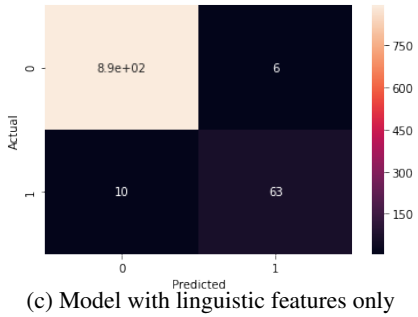
As arised in Section 4, one could question the impact of OOV terms (due to the ability to write French for example). After applying a simple spell checker (pyspellchecker) based on Levenshtein Distance (correcting unknown words of 4 length characters with an edit distance of 2 from an orginal word), we noted none improvement in predicting the certification with logistic regression using linguistic features for any MOOCs.

Figure 2: Confusion matrices corresponding to the predictions of logistic regression models for *ffl2* data

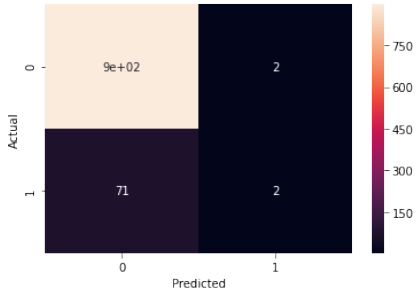
(a) Model with event-based features only



(b) Model with both event-based and linguistic features



(c) Model with linguistic features only



5.2 Second experiment: linguistically-based and event-based indicators over the time

Predicting learner success or failure with the available dataset after the MOOC has ended appears to be of little use in a practical context where such predictions would be needed before the MOOC ends. Furthermore, linguistic feature may have a greater importance in prediction if we limit the data to those available at the beginning of the MOOC, since fewer event-based indicators are available at that time. Therefore, we decided to create predictive models in the same way as in the previous experiment, with the same choices of features, but limiting the data to messages available at the end of the first week of the MOOC, then at the end of the second week, third week, etc. So we implemented two models per week (event-based indicators alone or coupled with linguistically-based indicators), with the models having access to more or less data depending on the week they corresponded to. We did not implement models based solely on linguistically-based features, on the assumption that if the model built with the data from the entire MOOC failed to make correct predictions, it would also fail to do so with less data. This experiment was performed on *ffl2*.

Figure 3 shows a graph plotting the evolution of the accuracy, precision, recall and F1-score obtained by the models over the weeks. The results of the models with and without linguistic features are merged because they are identical: again, the chosen linguistic features do not help the prediction. We observe that recall becomes good after three weeks of MOOC. This seems to correspond to the time needed to obtain enough information to predict the success or failure

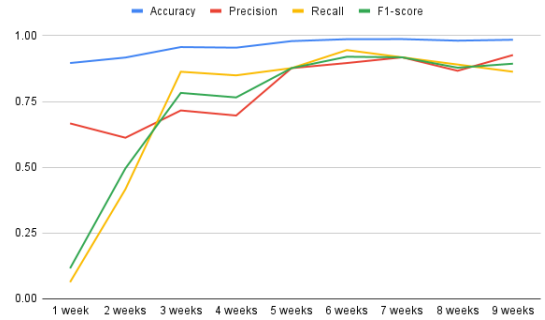
Table 4: Accuracy, precision, recall and F1-score for predictions obtained with *df2* data in three configurations

	Accuracy	Precision	Recall	F1-score
Model with event-based features only	0.97	0.95	0.97	0.96
Model with both event-based and linguistic features	0.97	0.95	0.97	0.96
Model with linguistic features only	0.67	0.78	0.20	0.32

Table 5: Accuracy, precision, recall and F1-score for predictions obtained with *ffl2* data in three configurations

	Accuracy	Precision	Recall	F1-score
Model with event-based features only	0.98	0.91	0.86	0.89
Model with both event-based and linguistic features	0.98	0.91	0.86	0.89
Model with linguistic features only	0.92	0.5	0.03	0.05

Figure 3: Evolution of accuracy, precision, recall, and F1-score of predictive models over time (number of weeks) of data available for *ffl2*



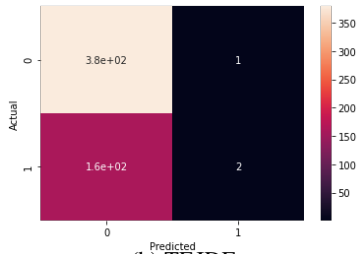
of the MOOC learners with the help of event-based features.

5.3 Third experiment: *TF.IDF* and word embeddings

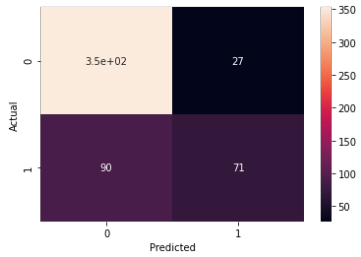
Our last experiment was to use non-specialized linguistically-based indicators as features for predictive models, in the hope that the language models used would detect elements or patterns in the text that would predict learner success or failure. We performed all these experiments on the corpus composed of a mix of *ffl1, ffl2* and *ffl3*. We chose to mix the datasets in order to have a larger amount of data, which allowed to create large training, test and validation sets (for the FastText and BERT models). Using as input features the values given by the *TF.IDF* method, we obtain the confusion matrix given in Figure 4b. Figures 5a and 5b give respectively the confusion matrices obtained on the test sets using the FastText and BERT models. Figure 4a gives for comparison the results obtained with the specific features as input, i.e. the same linguistic features as those used in the second experiment. The table 6 gives the calculations of accuracy, precision, recall and F1-score in each configuration.

Figure 4: Confusion matrix corresponding to the predictions of the classification models taking as input linguistically-based features and *TF.IDF* features

(a) Linguistically-based indicators related to the cognitive and affective engagement



(b) TF.IDF



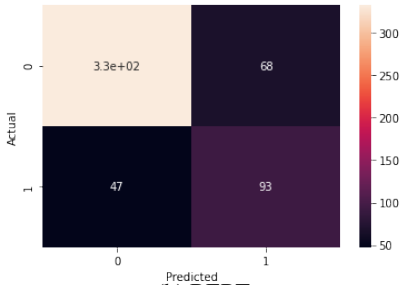
Features	A	P	R	F1-score
Ling.	0.70	0.67	0.01	0.02
TF.IDF	0.78	0.72	0.44	0.54
FastText	0.79	0.58	0.66	0.62
BERT	0.78	0.67	0.55	0.61

Table 6: Accuracy (A), Precision (P), Recall (R) and F1-score for predictions obtained with *ffl2* data for five types of features

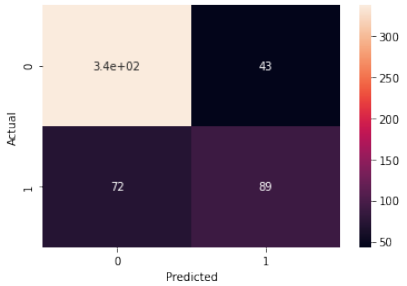
The model using linguistically-based features performs very poorly in prediction, with an F1-score close to 0. In the same way as in the first experiment, we observe that this model classifies almost all the items in the "failure" category (0). The model taking *TF.IDF* features as input makes more correct predictions of student success (true positives), but these predictions account for less than half of the successful

Figure 5: Confusion matrix corresponding to the predictions of the FastText and BERT word-embeddings based classification models

(a) FastText



(b) BERT



students, with the rest incorrectly predicted as failing.

Among these four models taking linguistically-based features only as input, the best prediction results are obtained by the models based on FastText and BERT word embeddings. These two models have very similar results, with slightly higher precision for BERT and, conversely, slightly higher recall for FastText. However, even though the results are better compared to the other models, they still perform poorly, with many misclassified items. This makes their contribution uninteresting compared to the models taking event-based indicators as features.

6 Discussion and perspectives

The state of the art on the notion of engagement and on its measurement in the framework of MOOCs showed the existence of several tracks, explored for the English language with sometimes contradictory results. We started by adapting some indicators to the French language in order to reproduce prediction experiments. To our knowledge, our work is one of the first studies with French speaking MOOCs. The results of our predictive models do not succeed to show an interesting contribution of lexicon-based approaches for measuring individual cognitive and affective engagement. However, this could be due to the fact that the chosen indicators were too simple or not very precise: it would be interesting to adapt more complex linguistic tools used for English to French, such as tools analyzing syntactic complexity or cohesion (Crossley et al., 2018). Linguistically-based approaches using deep representation gave better results compared to linguistically-based surface approaches. This track merits also to be explored in particular by searching how to make them more sensitive to the complexity and the abstraction of the analysed texts. Like Wen et al. (2014b), our results show that text processing may support global analyses but can hardly support individual follow-up. But this work requires to study other MOOCs and domains to confirm the observation. Deep learning approaches may change this conclusion. NLP techniques remain useful for providing additional information for building social network (Wise and Cui, 2018) or for fine-grained analysis such as dialogue acts analysis (Joksimovic et al., 2020).

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their valuable comments. This work was partially supported by the French *Agence Nationale de la Recherche*, within its *Programme d'Investissements d'Avenir*, with grant ANR-16-IDEX-0007.

REFERENCES

Atapattu, T., Thilakaratne, M., Vivian, R., and Falkner, K. (2019). Detecting cognitive engagement using word embeddings within an online teacher professional development community. *Computers & Education*, 140:103594.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Crossley, S., Ocumpaugh, J., Labrum, M. J., Bradfield, F., Dascalu, M., and Baker, R. (2018). Modeling Math Identity and Math Success through Sentiment Analysis and Linguistic Features. In *EDM*.

Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S., and Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, pages 6–14, New York, NY, USA. Association for Computing Machinery.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). A survey of the state of explainable ai for natural language processing. In *Proceedings of*

the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Ferri, P. (2019). Mooc, digital university teaching and learning analytics. opportunities and perspectives. *ITALIAN JOURNAL OF EDUCATIONAL RESEARCH*, page 13–26.

Fincham, E., Whitelock-Wainwright, A., Kovanović, V., Joksimović, S., van Staalduinen, J.-P., and Gašević, D. (2019). Counting Clicks is Not Enough: Validating a Theorized Model of Engagement in Learning Analytics. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 501–510, Tempe AZ USA. ACM.

Joksimovic, S., Jovanovic, J., Kovanovic, V., Gasevic, D., Milikic, N., Zouaq, A., and Van Staalduinen, J. (2020). Comprehensive analysis of discussion forum participation: from speech acts to discussion dynamics and course outcomes. *IEEE Transactions on Learning Technologies*, 13(1):38–51.

Joksimović, S., Poquet, O., Kovanović, V., Dowell, N., Mills, C., Gašević, D., Dawson, S., Graesser, A. C., and Brooks, C. (2018). How Do We Model Learning at Scale? A Systematic Review of Research on MOOCs. *Review of Educational Research*, 88(1):43–86. Publisher: American Educational Research Association.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Le, Q. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *International Conference on Machine Learning*, pages 1188–1196. PMLR. ISSN: 1938-7228.

Martinc, M., Pollak, S., and Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Moreno-Marcos, P. M., Muñoz-Merino, P. J., Maldonado-Mahauad, J., Pérez-Sanagustín, M., Alario-Hoyos, C., and Delgado Kloos, C. (2020). Temporal analysis for dropout prediction using self-regulated learning strategies in self-paced moocs. *Computers & Education*, 145:103728.

Piolat, A., Booth, R., Chung, C., Davids, M., and Pennebaker, J. (2011). La version française du dictionnaire pour le LIWC : modalités de construction et exemples d'utilisation. *Psychologie Francaise - PSYCHOL FR*, 56:145–159.

Ramesh, A., Goldwasser, D., Huang, B., III, H. D., and Getoor., L. (2013). Modeling learner engagement in moocs using probabilistic soft logic.

Reschly, A. L. and Christenson, S. L. (2012). Jingle, Jangle, and Conceptual Haziness: Evolution and Future Directions of the Engagement Construct. In Christenson, S. L., Reschly, A. L., and Wylie, C., editors, *Handbook of Research on Student Engagement*, pages 3–19. Springer US, Boston, MA.

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2020). Masked language model scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Tucker, C., Pursel, B., and Divinsky, A. (2014). Mining Student-Generated Textual Data In MOOCS And Quantifying Their Effects on Student Performance and Learning Outcomes Mining Student-Generated Textual Data in MOOCS and Quantifying Their Effects on Student Performance and Learning Outcomes. *Computers in Education Journal*, 5:84–95.

Wen, M., Yang, D., and Rosé, C. (2014a). Linguistic Reflections of Student Engagement in Massive Open Online Courses.

Wen, M., Yang, D., and Rosé, C. (2014b). Sentiment analysis in MOOC discussion forums: What does it tell us? *Proceedings of Educational Data Mining*, pages 130–137.

Whitehill, J., Williams, J., Lopez, G., Coleman, C., and Reich, J. (2015). Beyond Prediction: First Steps Toward Automatic Intervention in MOOC Student Stopout.

Wise, A. F. and Cui, Y. (2018). Learning communities in the crowd: Characteristics of content related interactions and social relationships in MOOC discussion forums. *Computers & Education*, 122(1):221–242. Publisher: Elsevier Ltd.

Yang, D., Wen, M., Howley, I., Kraut, R., and Rose, C. (2015). Exploring the Effect of Confusion in Discussion Forums of Massive Open Online Courses. pages 121–130.