Sentiment Analysis of Students' Reviews on Online Courses: A Transfer Learning Method

Thanh Vu Ngoc, Mai Nguyen Thi
School of Economics and Management
Hanoi University of Science and Technology
Hanoi, Vietnam
thanh.vn163755@sis.hust.edu.vn, mai.nt162625@sis.hust.edu.vn

Hang Nguyen Thi
School of Applied Mathematics and Informatics
Hanoi University of Science and Technology
Hanoi, Vietnam
hangnt.161384@sis.hust.edu.vn

Abstract

It is critical for higher education institutions to work on improvement of their teaching and learning strategy by examining feedback of students. Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing (NLP) and text analysis to systematically identify, extract and quantify states and subjective information. Transfer learning (TL) could be a ponder in machine learning centering on putting away information picked up whereas understanding one issue and applying it to a diverse but related issue. In this paper, we present the results from applying BERT, a transfer learning method for one of text classification problems. The model aims to predict state positive, negative, or neutral of an online course from students' reviews. This result will be compared with the contributions of authors Kastrati et al. (2020). The results which using BERT are quite good (Accuracy 88.93%), giving us more hope for future research.

Keyword:

Sentiment analysis, sentiment classification, online course, student feedback, MOOCs.

1. Introduction

Social media are giving the humus for the sharing of information and encounters and the development of community exercises (e.g., debating about different topics) (Dessì et al., 2019). As a result of this progression, an expanding number of course audits are being produced with the rise of Massive Open Online Courses (MOOCs), which offers teachers a chance to analyze, discover the opinions of learners and improve teaching strategies (Z. Liu et al., 2016). Working towards improving MOOCs, it is important to know students' opinions about the course and also the major course tools (Wen et al., 2014). That is the reason why Sentiment Analysis (SA) is widely applied to the materials such as reviews and survey responses by form, comment online about Cousera. SA is the computational treatment of opinions, sentiments and subjectivity of text (Medhat et al., 2014). These sentiments can be categorised either into two categories: positive and negative; or into an n-point scale, e.g., very good, good, satisfactory, bad, very bad (Prabowo & Thelwall, 2009).

Several different conventional machine learning algorithms such as SVM, Decision Tree, Naive Bayes (NB) and Boosting and 1D-CNN deep learning classifier using bag of words - BoW used for text classification (Kastrati, Arifaj, et al., 2020). In this paper, we use transfer learning methods for NLP, BERT (Bidirection al Encoder Representations from Transformers) for text classification task. The data set we use in this study are reviews of students who have directly studied online on the Coursera platform. Today, Coursera is a global online learning platform that offers anyone, anywhere, access to online courses and degrees from leading universities and companies. The results shows that the algorithm is suitable for the proposed data set.

In the first sections, we will conduct data set character analysis (data size, key characteristics), then we perform data preprocessing, removing, or replacing values, unsuitable characters. This is extremely important because if a word or character is misspelled, it can mislead the whole sentence. In the next section, we go into analysis of the model architecture: BERT. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers (Devlin et al., 2019). The rest of paper, we focus on interpreting results, drawing conclusions, and giving limitations. We also give some ideas for future research.

2. Literature Review

Sentiment analysis has been handled as a Natural Language Processing task at many levels of granularity (Agarwal et al., 2011). Sentiment analysis attempts to automatically identify and recognize opinions and emotions in text is represents a positive or negative or neutral opinion. This has been used in many areas such as business, politics, and especially in education to assess students'feedback. Recent research on social media use has demonstrated that sentiment analysis can reveal a variety of behavioral and affective trends (Wen et al., 2014). For example, Balahadia et al. (2016) developed a teacher's performanc evaluation tools using opinion mining with sentiment analysis. Forum messages in MOOCs (Massive Open Online Courses) are the foremost imperative source of data almost the social intuitive happening in these courses (Moreno-Marcos et al., 2018). Their study may help to identify strengths and weaknesses of faculty teachers based on students' feedback. There are a lot of research works about sentiment analysis published in recent years (Kastrati, Arifaj, et al., 2020) and there are a lot of previous papers that use algorithms in machine learning to analyze sentiment, such as: Lexicon-based (Rani & Kumar, 2017)-(Moreno-Marcos et al., 2018), NB, SVM (Nasim et al., 2017), Unsupervised methods (Dragoni et al., 2019), etc.

Lexicon based approach of sentiment analysis makes use of a sentiment lexicon to determine the polarity of a given textual content (Nasim et al., 2017). Rani & Kumar (2017) used a system that uses natural language processing in conjunction with NRC Emotion Lexicon to categorize emotions and sentiments. Sentiments are divided into two categories: positive and negative, and emotions are divided into eight categories: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, from which satisfaction or dissatisfaction is calculated. He argues that his SA system improves teaching and learning by performing transient emotional and sentiment analysis of multilingual student feedback on teacher performance and satisfaction. It is interesting to provide some nuances in terminology based on Munezero et al. (2014) and Mohammad & Turney (2013):

- **Emotion:** Pervasive among humans, and many are innate; the affective aspect of consciousness; a tate of feeling; a conscious mental reaction (as anger or fear) subjectively experienced as strong feeling usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body.
- **Sentiment:** An attitude, thought, or judgement prompted by feeling; a specific view or notion.

Moreno-Marcos et al. (2018) in their paper supposed that there are very few contributions that focus specifically on sentiment analysis on MOOCs (although there could be examples that use simples approaches for other purposes). They provided a number of different analysis of outcomes, which aim to explore and discuss patterns in learner behavior that can be helpful for further course enhancement. With data set related to Java programs, results show that among the supervised approaches, Random Forest provides the best results both for AUC and kappa.

In 2017, Nasim et al. described a sentiment analysis model trained using tf*idf and lexicon-based features to analyze the sentiments expressed by students in their textual feedback. The paper described a hybrid approach for performing sentiment analysis on student feedbacks: Random Forest and SVM (the dataset used in this paper comprises of 1230 comments extracted from our institutes educational portal). The hybrid model for sentiment analysis was trained using unigrams, bigrams, tf*idf and lexicon-based features.

In 2019, Cobos et al. provided a detailed description of the tool called edX-CAS ("Content Analyser System for edX MOOCs") - This tool was designed and developed at Universidad Autónoma of Madrid (UAM, Spain). They discovered promising solutions for knowledge adaptation to course material, but no resources that instructors can incorporate and use directly to implement SA in textual content have been discovered. That is why edX-CAS was introduced. Moreover, these analyses make use of multiple NLP techniques to conduct these processes, namely, tokenization, lemmatization, stopwords removal and POS-tagging.

In 2020, Lê et al. present the results from applying BERT. This model will also be applied to our paper; details about BERT will be provided later. They presented a transfer learning method that achieved high performance in Vietnamese dataset (The dataset used in this project is from the VLSP 2018 challenge).

The last two papers we collected are all using the data set on the Coursera.org. First, Wen et al. (2014) used three course: Accountable Talk: Conversation that works; Fantasy and Science Fiction: the human mind, our modern world; Learn to Program: The Fundamentals. They suggested Survival analysis: This is a statistical modeling technique used to model the effect of one or more indicator variables at a time point on the probability of an event occurring on the next time point. They see that within a specific course, the relationship between sentiment and dropout makes sense once one examines practices for expressing sentiment within that specific course context. Kastrati, Imran, et al. (2020) proposed in their paper a framework to automatically analyzing opinions of students expressed in reviews - Aspect-Based Sentiment Analysis (ABSA) framework. They said that ABSA is a sub-task of sentiment analysis which provides deeper understanding of the task at hand. Besides, the framework took advantage of the use of weak supervision strategy for prediction of the aspect categories that are critical factors in determining the effectiveness of online courses in general. However, this study only identifies some aspects related to the teaching performance of online courses.

Overall, the focus was on the accuracy and performance of the training data set to predict the best model. MLP and SVM are recognized as the outperforming models. Among them, BERT is introduced as a critical model for NLP based on transformer architecture. It is trained using two different tasks helping to learn the dependency between words in a sentence and between two sentences in a paragraph. Therefore, the characteristic of the language in which it was trained can be obtained. BERT also provided a pre-trained model in 104 languages (bert multilingual model) (Devlin et al., 2019).

3. Methods

3.1 Dataset

The dataset contains 21,940 reviews gathered from 15 courses on online leading platform Coursera (Kastrati, Imran, et al., 2020). Reviews are all in English language. Students'reviews are of different lengths, from 1 to 554 words, with an average of 25 words per review. Reviews are evaluated according to two parameters: the aspect category and aspect sentiment classification (positive, negative, and neutral). Specifically, the aspect categories include five dimensions:

1) ***Instructor*** - defined as reviews which related to instructor factors such as knowledge, skill and experience are commented on. For example, "Excellent course! Mr. Yakov Chaikin is a great instructor. Many thanks to him!!"

2) ***Content*** - refers to learners' responses to topics of the course including learning material, projects, such as This course was a great intro to these concepts and helpful guide to getting things set up and getting used to the mooc format, as well! a few times it seemed like the slides jumped right in while skipping over a bit of context, but was able to orient myself with some googling and asking friends some basic questions to figure things out.."

3) **Structure** - according to Devlin et al. (2019), it is devoted to reviews dealing with structural aspects of the course such as structure of modules and learning objectives. For instance, "The course design and assignment are consistent with each other. it not only helps a beginner avoid many unnecessary works, focusing on more relative knowledge, but also leave enough information for cs-like major to explore some other details. the course must take a lot of effort and heart of teaching teams to make this excellent."

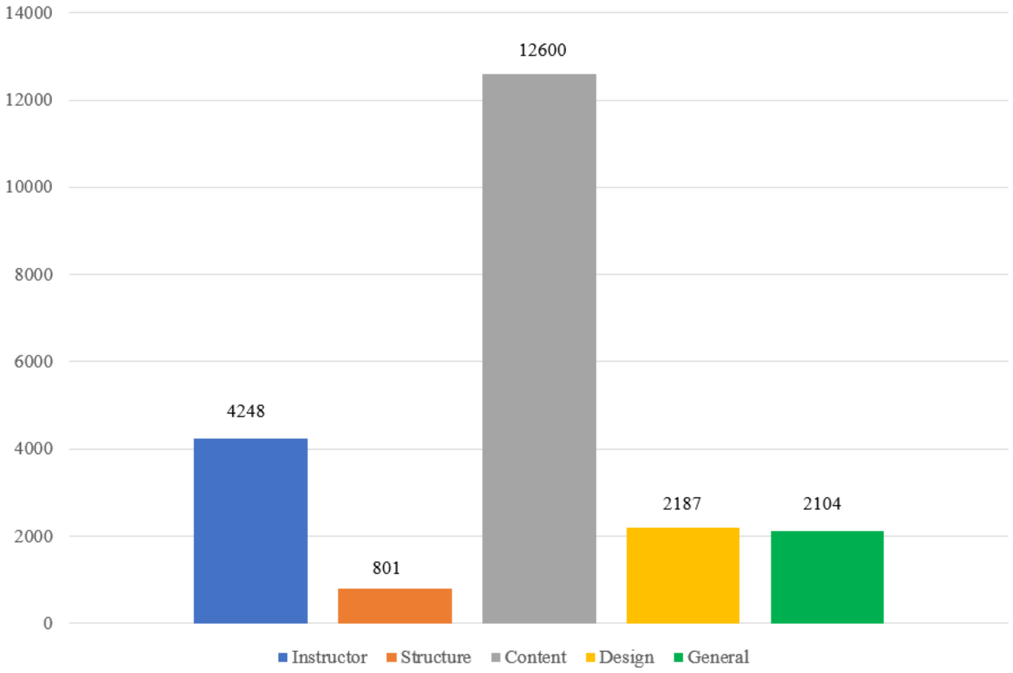

Figure1. Number of reviews with five course-related dimensions

4) ***Design*** - relates to reviews which students have commented on both course content and course structure. For example, "Well organized, well presented and has adequate information to introduce you to the capabilities of the sqlite."

5) ***General*** - includes student feedback on all aspects of the course, general feelings after going through the course. For instance, "Excellent course, charismatic teacher. I really enjoyed this course and hope i can move onto the specialisation. Thank you."

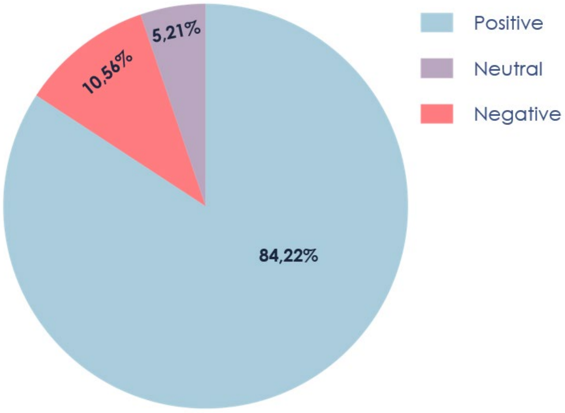

Figure 2. Percentage of sentiment in the dataset

3.2 Preprocessing

Data from the real world is never perfect (Famili et al., 1997). Preprocessing of data is an essential activity which will help to improve the quality of the data and successively the mining results (Sudheer Reddy et al., 2013). In particular, for on-line environment, the document is not often in formal writing. Especially for teenagers, they prefer using a great number of emojis, shortened forms of words, special symbols and characters, misspelling words, grammar mistakes or conjoined words.

Before being included in models, the data should follow preprocessing steps:

- **Step 1:** All characters indicating money value such as “100k” or “100,000$” should be replaced by words “price money” (money value). Also, characters indicating percentage value such as ”50%” should be replaced by word “Percentage”.
- **Step 2:** All special characters as “=”, “?”, “#”, “+”, “-”, “$”, “”, “&” And so on, and emojis should be removed.
- **Step 3:** Shortened forms of word.
- **Step 4:** This is the most important step. The need-tosolved problem is a multi-label classification, so a great number of potentially predicted labels for a review can lead to a confusion for the model. Therefore, we separated long reviews into many sentences, and then relabeled add. These relabels are definitely included in review labels.

3.3 Model Architecture

BERT (Bidirectional Encoder Representations from Transformers), which leverages a multi-layer multi-head self-attention (called transformer) together with a positional word embedding, is one of the most successful deep neural network model for text classification in the past years (Lu et al., 2020). BERT has inspired many recent NLP architectures, training approaches and language models, such as Google’s TransformerXL, OpenAI’s GPT-2, XLNet, ERNIE2.0, RoBERTa, etc (www.analyticsvidhya.com, 2019). The transformer-based pre-trained language model BERT has helped to improve state-of-the-art performance on many natural language processing (NLP) tasks (de Vries et al., 2019). Instead of conducting directional modeling of context of a word, transformers like BERT model relations between all pairs of tokens using a self-supervised strategy (D. Liu & Miller, 2020). BERT is a deep bidirectional representation model for general-purpose “language understanding” that learns information from left to right and from right to left. A pre-trained BERT model based on 800M words from BooksCorpus and 2,500M words from English Wikipedia is made available (Lu et al., 2020). BERT has two models:

- BERT-base: 12 Encoders with 12 bidirectional selfattention heads.
- BERT-large: 24 Encoders with 24 bidirectional selfattention heads.

We used Bert-base-uncased for this project. BERT takes an input of a sequence of no more than 512 tokens and outputs the representation of the sequence (C. Sun et al., 2019). Bert architecture have 12 Encoders with 12 bidirectional self-attention heads and where the text has been lowercased before WordPiece tokenization. e.g., “James Bond” becomes “james bond”. It also removes accent markers.

Because this is a sentence classification task, we ignore all except the first vector (the one associated with the [CLS] token). The one vector we pass as the input to the softmax regression model. In this paper, we looked for the proper methods with three steps. The input sentence is a sequence of n words: x = ([CLS], x1, x2, ..., xn, [SEP]), where [CLS] token stands for “Classification” and will become apparent later on and [SEP] stands for “Separate” is a dummy token not used for the task. Set BERT(.) as a pre-trained BERT model simulation function. Sentence x after being tokenized and passed to the BERT() function becomes hidden representation h = BERT(x), h has the size rh ∗ |x| where rh is the size of the hidden dimension and |x| is the length of the input sentence. We use h[CLS] - the representation of [CLS], which is review representation of the whole input sentence. We concatenation h[CLS] of 4 last layer then we obtain h[CLS] summary subsequently passed through a dense layer followed by a softmax function l = softmax(W ∗ h[CLS] + b), where W belongs to Rm∗rh , where m is the number of class.

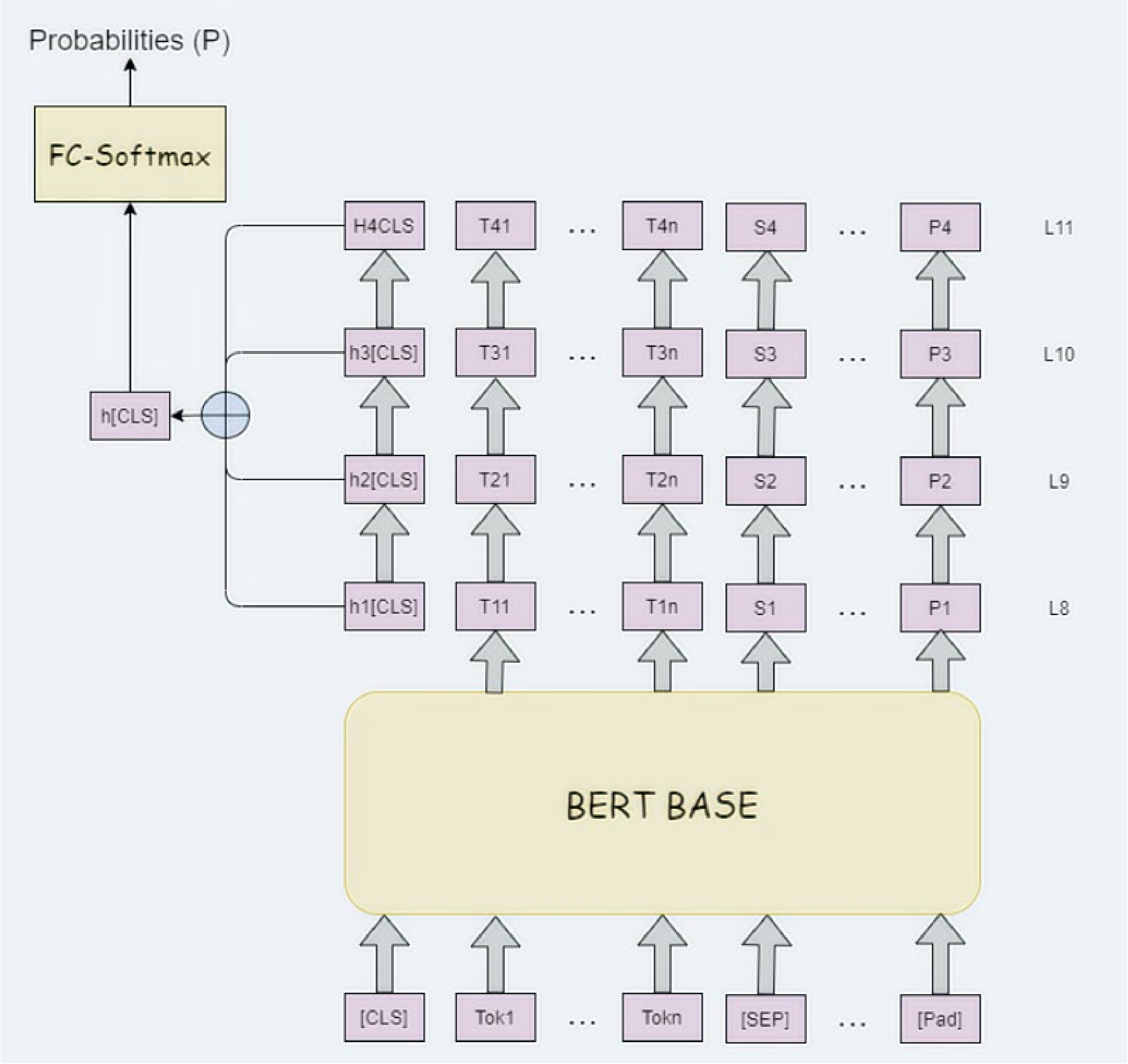

Figure 3. The steps of BERT

The first step is to use the BERT tokenizer to first split the word into tokens. Then, we add the special tokens needed for sentence classifications (these are [CLS] at the first position, and [SEP] at the end of the sentence). The last step the tokenizer does is to replace each token with its id from the embedding table which is a component we get with the trained model. Read The Illustrated Word2vec for a background on word embeddings.

4. Results and Discussion

In this study, we use python and the powerful library scikit learn to calculate the coefficients Precision (P), Recall (R) and F1 score. First, remember that true positive (TP) are samples that were classified positive and are really positive. False positive samples (FP) are samples that were classified positive but should have been classified negative. Similarly, false negative (FN) were classified negative but should be positive. Here, TP, FP and FN stand for the respective number of samples in each of the classes.

Precision is defined as the ratio of true positive scores among those classified as positive (TP + FP). Recall is defined as the ratio of the number of true positives among those that are actually positive (TP + FN). When Precision = 1, all scores found are actually positive, meaning there are no negative points mixed into the results. However, Precision = 1 does not guarantee that the model is good, as the question is whether the model has found all the positives. If a model finds only one positive point for which it is most certain, then we cannot call it a good model. When Recall = 1, every positive is found. However, this quantity does not measure how many negatives are mixed in. If the model classifies every point as positive, then surely Recall = 1, but it is easy to see that this is a very bad model. F1 score =2(P∗R)/P+R. This is just the weighted average between precision and recall. The higher precision and recall are, the higher the F1 score is. You can directly see from this formula, that if P=R, then F1=P=R

4.1 Aspect Categories

The obtained performance of the BERT's technique using positional word embedding with respect to Precision, Recall and F1 score is depicted in **Table 1**. We also compare with the results of Kastrati, Arifaj, et al. (2020). They used different algorithms to test the accuracy such as Decision Tree, SVM using tf*idf and tf.

Table 1: Performance of techniques (aspect categories)

Technique	Precision (%)	Recall (%)	F1 (%)
BERT	82.68	82.68	82.68
Decision Tree using tf*idf	78.81	78.86	78.83
SVM using tf*idf	76.11	75.92	76.01
Decision Tree using tf	62.65	61.82	61.04
SVM using tf	61.08	60.68	60.88

The BERT technique shows the F1 score is 82.68%. Meanwhile the second-best result is 78.83%, which is 3.85% bigger. This can be assumed that BERT will be a new breakthrough in natural language processing technology. The reason for this result may be that BERT processes two-dimensional contexts, exploiting the meaning of sentences more clearly.

4.2 Aspect Sentiment Classification

Based on the labels distinguish the sentiments: positive, negative, and neutral. The BERT technique is also more efficient than other algorithms implemented by the authors Kastrati, Arifaj, et al. (2020). The specific results are presented in **Table 2**.

Table 2: Performance of techniques (aspect sentiment classification)

Technique	Precision (%)	Recall (%)	F1 (%)
BERT	88.98	88.92	88.94
Decision Tree using tf*idf	88.72	88.61	88.67
SVM using tf*idf	88.46	88.15	88.31
Decision Tree using tf	78.86	76.96	77.90
SVM using tf	76.94	75.93	76.43

The BERT technique shows quite good results with F1 score is 88.94%. Besides, decision tree and SVM using tf*idf algorithms also work very well when giving high results (88.67% and 88.31%). the results are much higher than model for prediction of the aspect categories. This will be explained by the truth that there's less covering between extremity labels than between aspect categories (Kastrati, Arifaj, et al., 2020).

5. Conclusion

In this article, we have exploited that BERT algorithm which using to analyze the sentiment of students who taking an online course on Coursera. The results show that the algorithm is quite suitable for the data set. We believe that the same effect can be applied to other datasets, which can be investigated further. Compared with other studies along with the implementation process, we have limitation: Although the obtained results are superior to some tests of other papers, the problem of multi-label reviews has not been solved. Besides, the results are not too different from other techniques. We have to find ways to combine and improve more.

In the future, as mentioned, we will experiment with more datasets of different fields to explore more about BERT as well as expand to the multi-label problem.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment Analysis of Twitter Data. *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 30–38. https://www.aclweb.org/anthology/W11-0705

Balahadia, F. F., Fernando, Ma. C. G., & Juanatas, I. C. (2016). Teacher's performance evaluation tool using opinion mining with sentiment analysis. *2016 IEEE Region 10 Symposium (TENSYMP)*, 95–98. https://doi.org/10.1109/TENCONSpring.2016.7519384

Cobos, R., Jurado, F., & Blázquez-Herranz, A. (2019). A Content Analysis System That Supports Sentiment Analysis for Subjectivity and Polarity Detection in Online Courses. *IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje*, *14*(4), 177–187. https://doi.org/10.1109/RITA.2019.2952298

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). BERTje: A Dutch BERT Model. *ArXiv:1912.09582 [Cs]*. http://arxiv.org/abs/1912.09582

Dessì, D., Dragoni, M., Fenu, G., Marras, M., & Recupero, D. R. (2019). Evaluating neural word embeddings created from online course reviews for sentiment analysis. *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 2124–2127. https://doi.org/10.1145/3297280.3297620

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Dragoni, M., Federici, M., & Rexha, A. (2019). An unsupervised aspect extraction strategy for monitoring real-time reviews stream. *Information Processing & Management*, *56*(3), 1103–1118. https://doi.org/10.1016/j.ipm.2018.04.010

Famili, A., Shen, W.-M., Weber, R., & Simoudis, E. (1997). Data Preprocessing and Intelligent Data Analysis. *Intelligent Data Analysis*, *1*(1), 3–23. https://doi.org/10.3233/IDA-1997-1102

Kastrati, Z., Arifaj, B., Lubishtani, A., Gashi, F., & Nishliu, E. (2020). Aspect-Based Opinion Mining of Students' Reviews on Online Courses. *Proceedings of the 2020 6th International Conference on Computing and Artificial Intelligence*, 510–514. https://doi.org/10.1145/3404555.3404633

Kastrati, Z., Imran, A. S., & Kurti, A. (2020). Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs. *IEEE Access*, *8*, 106799–106810. https://doi.org/10.1109/ACCESS.2020.3000739

Lê, N. C., The Lam, N., Nguyen, S. H., & Thanh Nguyen, D. (2020). On Vietnamese Sentiment Analysis: A Transfer Learning Method. *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 1–5. https://doi.org/10.1109/RIVF48685.2020.9140757

Liu, D., & Miller, T. (2020). Federated pretraining and fine tuning of BERT using clinical notes from multiple silos. *ArXiv:2002.08562 [Cs]*. http://arxiv.org/abs/2002.08562

Liu, Z., Liu, S., Liu, L., Sun, J., Peng, X., & Wang, T. (2016). Sentiment recognition of online course reviews using multi-swarm optimization-based selected features. *Neurocomputing*, *185*, 11–20. https://doi.org/10.1016/j.neucom.2015.12.036

Lu, Z., Du, P., & Nie, J.-Y. (2020). VGCN-BERT: Augmenting BERT with Graph Embedding for Text Classification. In J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, & F. Martins (Eds.), *Advances in Information Retrieval* (pp. 369–382). Springer International Publishing. https://doi.org/10.1007/978-3-030-45439-5_25

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, *29*(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Moreno-Marcos, P. M., Alario-Hoyos, C., Muñoz-Merino, P. J., Estévez-Ayres, I., & Kloos, C. D. (2018). Sentiment analysis in MOOCs: A case study. *2018 IEEE Global Engineering Education Conference (EDUCON)*, 1489–1496. https://doi.org/10.1109/EDUCON.2018.8363409

Munezero, M., Montero, C. S., Sutinen, E., & Pajunen, J. (2014). Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions on Affective Computing*, *5*(2), 101–111. https://doi.org/10.1109/TAFFC.2014.2317187

Nasim, Z., Rajput, Q., & Haider, S. (2017). Sentiment analysis of student feedback using machine learning and lexicon based approaches. *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)*, 1–6. https://doi.org/10.1109/ICRIIS.2017.8002475

Prabowo, R., & Thelwall, M. (2009). Sentiment analysis: A combined approach. *Journal of Informetrics*, *3*(2), 143–157. https://doi.org/10.1016/j.joi.2009.01.003

Rani, S., & Kumar, P. (2017). A Sentiment Analysis System to Improve Teaching and Learning. *Computer*, *50*(5), 36–43. https://doi.org/10.1109/MC.2017.133

Sudheer Reddy, K., Kantha Reddy, M., & Sitaramulu, V. (2013). An effective data preprocessing method for Web Usage Mining. *2013 International Conference on Information Communication and Embedded Systems (ICICES)*, 7–10. https://doi.org/10.1109/ICICES.2013.6508197

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), *Chinese Computational Linguistics* (pp. 194–206). Springer International Publishing. https://doi.org/10.1007/978-3-030-32381-3_16

Wen, M., Yang, D., & Rosé, C. P. (2014). *Sentiment Analysis in MOOC Discussion Forums: What does it tell us?*

www.analyticsvidhya.com. (2019, September 25). What is BERT | BERT For Text Classification. *Analytics Vidhya*. https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/