



Students feedback analysis model using deep learning-based method and linguistic knowledge for intelligent educational systems

Asad Abdi^{1,2} · Gayane Sedrakyan^{2,3} · Bernard Veldkamp³ · Jos van Hillegersberg² · Stéphanie M. van den Berg³

Accepted: 11 February 2023 / Published online: 13 March 2023
© The Author(s) 2023

Abstract

Student feedback analysis is time-consuming and laborious work if it is handled manually. This study explores the use of a new deep learning-based method to design a more accurate automated system for analysing students' feedback (called DTLP: *deep learning and teaching process*). The DTLP employs convolutional neural networks (CNNs), bidirectional LSTM (BiLSTM), and attention mechanism.

To the best of our knowledge, a deep learning-based method using a unified feature set, which is representative of word embedding, sentiment knowledge, sentiment shifter rules, linguistic and statistical knowledge, has not been thoroughly studied with regard to sentiment analysis of student feedback. Furthermore, DTLP uses multiple strategies to overcome the following drawbacks: *contextual polarity; sentence types; words with similar semantic context but opposite sentiment polarity; word coverage limit of an individual lexicon; and word sense variations*. To evaluate the DTLP, we conducted an experiment on a large volume of students' feedback. The results showed (i) DTLP outperforms the existing systems in the field, (ii) DTLP that learns from this unified feature set can acquire significantly higher performance than one that learns from a feature subset, (iii) the ensemble of sentiment shifter rules, word embedding, statistical, linguistic, and sentiment knowledge allows DTLP to obtain significant performance, and (iv) an attention mechanism into CNN-BiLSTM improves the performance of DTLP. In addition, the deployed method looks for potential causes behind student feedback.

Keywords Student feedback · Deep learning · Sentiment analysis · Educational data mining · Intelligent learning environments

1 Introduction

Technological advancement has innovated education dramatically. Among multiple ways of contributing to teaching and learning processes, educational institutions use

technology as an instrument for gathering information about student experiences and assessing/adjusting their teaching approaches. Educational institutions should provide a healthy learning environment and must create and manage learning content to support and facilitate a successful teaching and learning process. In addition, for educational

✉ Asad Abdi
A.abdi@Derby.ac.uk; s.abdiesfandani@utwente.nl

Gayane Sedrakyan
g.sedrakyan@utwente.nl

Bernard Veldkamp
b.p.veldkamp@utwente.nl

Jos van Hillegersberg
j.vanhillegersberg@utwente.nl

Stéphanie M. van den Berg
stephanie.vandenberg@utwente.nl

¹ Department of Computing and Mathematics Faculty of Science and Engineering, University of Derby, Derby, UK

² Department of Industrial Engineering and Business Information Systems, University of Twente, Enschede, The Netherlands

³ Department of Cognition, Data, and Education, University of Twente, Enschede, The Netherlands

institutions, student achievement is critical. The level of satisfaction demonstrates a clear understanding of the university environment, and the services offered to students. Student satisfaction serves as a method to assess the quality of education achieved and the efficiency of the institution's operations. Higher education institutions are required to change their programmes, procedures, and structures in accordance with students' needs.

Students can share their ideas through a variety of methods, including classroom feedback, clickers, mobile phones, and social media such as Facebook and Twitter. Students' feedback aids instructors in identifying their strengths and shortcomings so that they may make necessary modifications to their teaching and curriculum to improve student learning quality. Curriculum design, learner behaviour, and faculty management decisions are also influenced by the feedback received from students and faculty members. Furthermore, student feedback can be also used to enhance the course content indefinitely as well as delivery for eLearning. In the learning and evaluation process, feedback is a highly effective tool. In earlier systems, students had to give feedback on paper. After filling out feedback forms, these paper forms were collected by the faculties and class teachers. Then, they calculated the overall grade for each subject and each teacher or faculty. Nowadays, students also use social media to express their opinions and feelings. Subsequently, a teacher has to read all the feedback. However, this process is still time-consuming and also costs a lot of resources to conduct this analysis either using paper or online feedback forms and/or social media.

In educational institutions, it is usually required that at the end of each term, students are invited to complete a questionnaire that is designed to collect students' perceptions and feedback on their learning experience in the course and the facilities of the university in terms of the classroom, teaching quality, improvement needs, library, etc. These questionnaires usually consist of Likert-scale questions, Yes/No, open-ended qualitative questions, or a mix. In this research, we also use the outcome of an online survey that we conducted at the University of Twente to collect students' feedback regarding the various aspects of learning experiences in a bachelor-level module. Through such surveys, the university aims to improve students learning and teaching quality. Furthermore, the main goal of this type of exercise (e.g. "questionnaire") is to allow instructors and senior managers to evaluate comments and enhance students' learning experiences and processes.

In general, much attention is paid to summarize quantitative feedback (e.g. *the statistical comparisons are calculated, analysed, and presented*), while qualitative feedback from students is not fully taken into account, as manual analysis of each student's comments can be extremely

difficult, due to the large-scale feedback. However, analysing the qualitative feedback data can provide valuable insights into teaching practices, course content, and curriculum. Handling students' opinions expressed in reviews is a quite difficult, time-consuming, and tedious task as it is typically performed manually by human intervention. While such a task may seem easy for small-scale courses that involve just a few students, it is impractical for large-scale cases. On the other hand, most educational institutions have the intent of collecting a large amount of data. However, it is important to extract opinions from huge amounts of opinionated content and present them in a simple comprehensible format. We expect that a sentiment analysis algorithm would be able to help users with this issue. Sentiment analysis is an application in natural language processing. Its main task is to analyse and predict the users' subjective sentimental polarities towards events or items.

In our study, we ask the following question: '*Can DTLP analyse students' opinions expressed in review or comment?*' For this purpose, our hypotheses include (i) a combined feature set constructed using feature vectors based on word embedding, sentiment knowledge, linguistic and statistical knowledge can potentially improve the performance of classification, and (ii) a joint network with CNN, BiLSTM, attention layer, and fully connected layer is capable to generate a significant performance.

Problem statement: Although researchers proposed different methods for sentiment analysis of students' feedback, these approaches have not yet matured to the point where they can solve existing issues, so there is still space and opportunity for development and improvement. In this research, we focus on sentiment analysis in students' feedback. Our objective is twofold. Firstly, we present a new deep learning-based approach to classify a student report, observation, or textual expression of an individual's viewpoint into a "*positive*" or "*negative*" context. DTLP is based on the CNN, BiLSTM algorithm, and attention mechanism for classification at the sentence level. DTLP (Fig. 1) receives word-level features as input through CNN, then an encoded feature from BiLSTM is fed into the attention layer. Subsequently, the concatenation of the sentence-level feature and the output of the attention layer is taken into account as the sentence-level vector representation (SLVR). Consequently, the SLVR is passed into the fully connected layer to reveal the feedback label (*positive/negative*). In this research work, we present a novel method that integrates the word embedding approach, sentiment shifter rules, statistical, linguistic, and sentiment knowledge to solve student feedback sentiment analysis problems. In addition, we propose a joint network with a CNN, BiLSTM, and attention layer to obtain the final sentiment classification results.

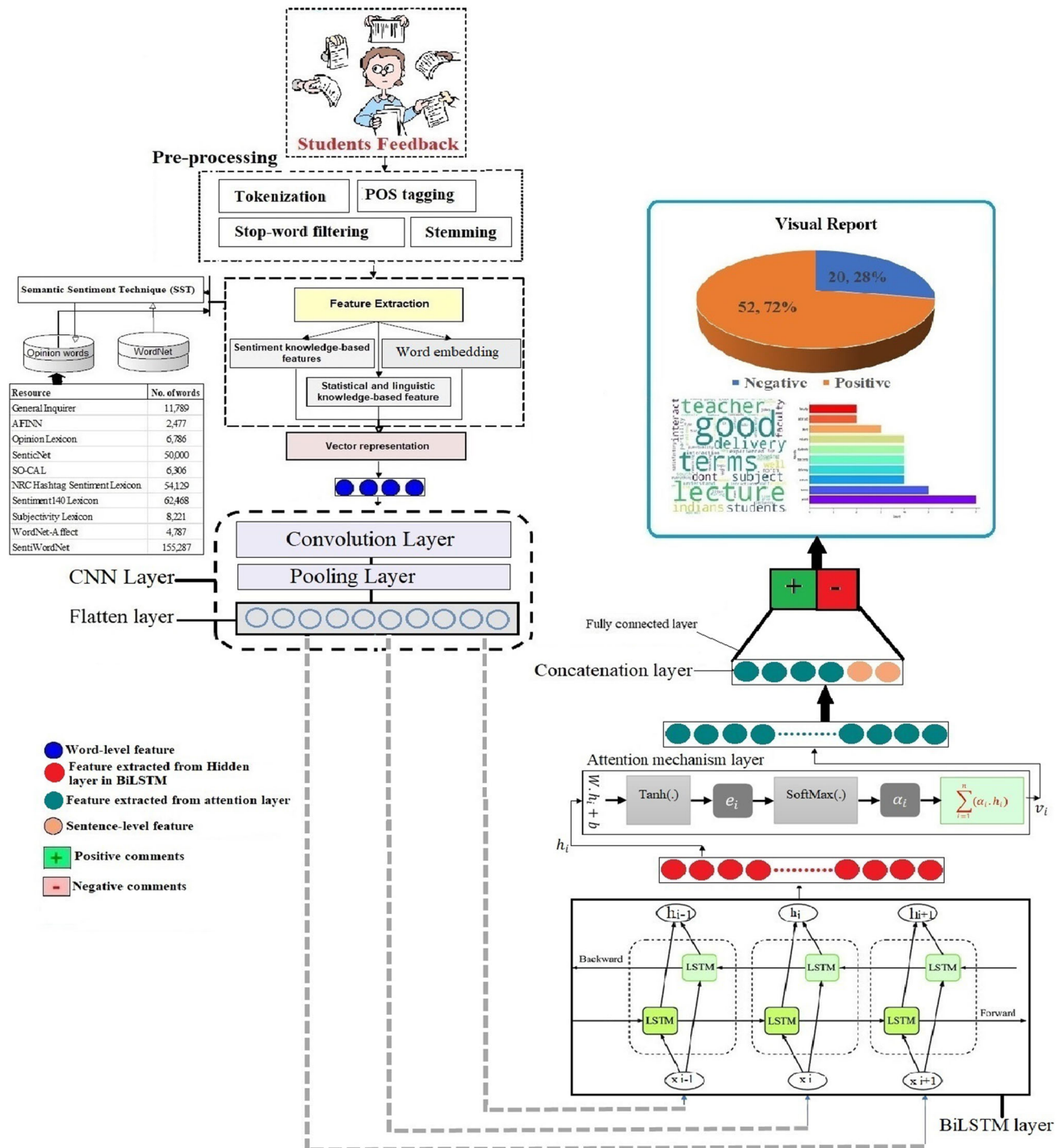


Fig. 1 The architecture of the DTLP

Secondly, we design and develop the method (DTLP) and present an experimental study to determine how well the DTLP performs in comparison with the other current systems.

The DTLP employs different strategies to overcome the following shortcomings that may be summed up as follows:

- (1) Words with the same semantic context but completely opposite emotional polarities: Several proposed methods used the word embedding approach (e.g. word2vec) to represent a vector of each word. A word embedding model does not consider the sentiment polarity of words (Araque et al. 2017). Consequently, two words with different polarities can be mapped

into close vectors (e.g. “like” and “dislike”). Therefore, a word embedding-based vector representation does not include enough sentiment information and is unable to accurately obtain the sentiment of a message. To address this issue, the word embedding approach can be enriched by the prior sentiment knowledge (PSK). PSK consists of complementary information that does not exist in the word embedding model. A sentiment lexicon can be considered a PSK.

- (2) **Word sense variations:** The word embedding approach is not able to distinguish a word’s sense and make a single representation per word form. For instance, the word vector for “apple” as a company or as a fruit is similar. Thus, DTLP also applies a strategy to solve this issue.
- (3) **Limited word coverage of an individual lexicon:** Several dictionaries are combined since they complement each other. Integrating different dictionaries can overcome the limitations of word coverage.
- (4) **Contextual polarity (*context-based sentiment analysis*):** It means that the polarity of a word is changed based on the context. Contextual polarity indicates *i*) the negation, which can appear in a different place in a message (for example, given the message “*the course was not attractive*”, the polarity of the sentence is changed due to the negation word “not”); *ii*) specific particles such as “*despite*” and “*but*” affecting the sentiment of a message (as an example, given the message “*the course material was good, but the teacher’s knowledge was not enough*”). The polarity of the phrase “*the course material was good*” is changed by the “*but*” clause.
- (5) **Sentence types:** In addition, the types of a sentence (e.g. *subjective, comparative, conditional/question, sarcastic*, etc.) (Chen et al. 2017) can also affect the sentiment analysis. Most of the existing research did not consider the sentence type.
- (6) **Different traditional machine learning-based methods (TMLBMs) have been proposed by researchers for this kind of task.** However, these methods fail to provide significant results because *i*) the feature engineering must be done by an expert, as the performance of a machine learning-based approach depends on how precisely the features are extracted and selected, which is also a quite difficult and time-consuming process; however, unlike the machine learning-based approach, a deep learning-based algorithm can automatically extract features and learn from the data and *ii*) a machine learning-based approach also has a limitation when it comes to capturing the dependence between words since it implicitly assumes that the words are independent.

On the other hand, a deep learning-based approach, for example, BiLSTM-based architecture, can handle sequential processes (*considers word order/syntactic structure*), consider long-range dependencies between tweet words, and guarantee better performance compared to other methods.

- (7) **Student feedback modelling:** Both traditional machine learning-based methods and a deep learning-based method (DLBM) need pre-processing of the text into numerical feature vectors. Most of the existing methods use the bag-of-words (BOW) approach for text representation and information retrieval. However, the BOW method has two major shortcomings: It leads to sparsity, i.e. many features have a zero value, and a high-dimensional feature vector due to the large size of the vocabulary. Also, the co-occurrence statistics between words are not taken into account. An N-gram is also a popular approach that, unlike the BOW method, considers the word order. However, it also suffers from high dimensionality and sparsity in data. To address these problems, a word embedding model for vector representation was employed. The statistical, linguistic, and sentiment knowledge features are considered in addition to the word embedding features to enhance the feature vectors.

Contributions: Our significant contribution in this paper can be summarized as follows:

- (1) Our work is a new deep learning-based method in which a unified feature set which is representative of word embedding, sentiment knowledge, sentiment shifter rules, statistical and linguistic knowledge is employed to solve the classification problem.
- (2) We proposed a joint network using the CNN, BiLSTM, and attention mechanism to take advantage of obtaining coarse-grained local features, sequential processing, to emphasize meaningful features and learn the different weights of words.
- (3) We employed several techniques/strategies to solve various challenges: *(i)* sentiment shifter rules; *(ii)* integration of word embeddings and sentiment information; *(iii)* types of sentences; *(iv)* semantic relationships between words and information about word order; *(v)* word sense variations; *(vi)* sentiment score calculation; and *(vii)* contextual polarity. This helps our method in achieving superior performance.
- (4) The DTLP employs several types of resource-information latent in a sentence to *(i)* learn a better sentence representation; *(ii)* create augmented vector; and *(iii)* obtain considerable performance. A hybrid vector is created to represent each sentence using the

statistical and linguistic knowledge-based, sentiment-based, and word embedding-based.

- (5) The DTLP combines various sentiment lexicons which complement each other to overcome the word coverage limitation.
- (6) We conducted extensive experiments for performance evaluation and comparison. We report our results on the real-world dataset. Furthermore, in addition to the existing dataset, we have also conducted an online survey to collect students' feedback regarding the various issues mentioned in the application form as survey questions. The results demonstrate that the DTLP obtains significant performance on all the measure metrics and confirms the suitability of our proposed method.
- (7) DTLP not only is able to handle a large amount of students' feedback, but also it can help to improve the quality of teaching, teacher's knowledge, and performance in the education system by implementing fine-grained aspect-based sentiment analysis in student feedback.

Paper organization: The rest of the paper is organized as follows. Section 2 discusses the recent related study as well as a brief overview of sentiment analysis. Section 3 presents our research methodology. Section 4 introduces our proposed method, DTLP. Next, in Sect. 5, the experimental results are reported. Finally, the conclusion and future work are presented in Sect. 6.

2 Literature review

The discovery of hidden knowledge on students' feedback data is essential to provide educational institutions and teachers with important insights that can be used in their daily practices.

The importance of student feedback in the teaching and learning process: The efficacy of feedback is determined by how the feedback data are processed and used. To enhance the quality of education in several areas such as education, teaching materials, and evaluation, all institutions gather various types of feedback. Effective utilization of student feedback can improve instruction and the curriculum, which, in turn, will lead to improved learning experiences and student outcomes earlier during a learning process (Sedrakyan 2016; Derick et al. 2017).

Learning analytics is a research area for the study of educational data mining to explore ideas and provide assistance in the decision-making process. Specifically, sentiment analysis is a procedure that can be used for qualitative examination of student comments at different levels such as school and university. Sentiment analysis is a

method to identify and classify the sentiment of a text into a positive or negative category using natural language processing and text analysis. In other words, sentiment analysis is the task of identifying users' emotions, attitudes, and opinions about any particular person, subject, or field. The sentiment analysis process includes the following steps: data collection (*the collection of sentiment text or user's opinion in reviews*), data preparation (*pre-processing, data cleaning, or removing irrelevant terms*), review analysis technique (*e.g. machine learning methods*), and sentiment classification (*e.g. positive or negative*). Sentiment analysis may be done at three levels: document, sentence, and aspect. Furthermore, it may also be divided based on the methods utilized into lexicon-based sentiment analysis, machine learning-based sentiment analysis, and the hybrid approach.

Sentiment analysis has lately gained favour in the field of education. In the field of education, sentiment analysis is performed during the assessment process to investigate hidden knowledge and opinions in open-ended questions. Researchers became more interested in how to improve students' learning experience by incorporating student feedback into the teaching and learning activities. The majority of research works concentrated on analysing quantitative data. However, some works on qualitative data were performed using sentiment analysis. For instance, Kaewyong, et al. (Kaewyong et al. 2015) presented a lexicon-based technique for document-level orientation evaluation on students' comments to evaluate instructors. Nasim, et al. (Nasim et al. 2017) presented a hybrid approach including a machine learning-based approach and lexicon-based method for analysing the sentiment orientation of students' written comments on a course. Furthermore, Misuraca, et al. (Misuraca et al. 2021) presented a lexicon-based method to determine the sentiment orientation of student feedback.

Table 1 presents the summary and comparisons of the features of the existing methods (✓ indicates their strengths and —their shortcomings). The literature revealed that several methods have been proposed for students' feedback analysis but these methods still have some drawbacks that need to be addressed. Furthermore, the table shows that there is not a method that integrates several factors to assess students' feedback.

Summing up feedback from students is always a valuable source of knowing aspects related to the learning process quality, classroom conditions, and teacher performance, teachers can enhance their instructional techniques by studying and comprehending the student's situation. Hence, different systems have been proposed to explore students' feedback in order to enhance learning outcomes. However, with respect to the DTLP, most of the existing methods used shallow NLP techniques. A few works propose a method

Table 1 A comparison of DTLP with other related research work

System [Ref]	Year	Approach	Word embedding	Word sense disambiguation	Sentiment-encoded word embedding	Sentence types	Sequential processing	Sentiment dictionary combination	Sentence types	Contextual polarity	
										sentiment shifter rules	Negation handling
DTLP	–	CNN, BiLSTM, attention layer, and linguistic knowledge	✓	✓	✓	✓	✓	✓	✓	✓	✓
Misuraca et al. (2021)	2021	Lexical-based	–	–	–	–	–	–	–	–	✓
Tamrakar et al. (2102)	2021	SVM, BOW, and TF-IDF	–	–	–	–	–	–	–	–	–
Katragadda et al. (2020)	2020	ANN ^a	–	–	–	–	–	–	–	–	✓
Lwin et al. (2020)	2020	SVM and TF ^b	–	–	–	–	–	–	–	–	–
Failed (2019a)	2019	CNN and SVM	✓	–	–	–	–	–	–	–	–
Failed (2019b)	2019	Lexical-based	–	–	–	–	–	–	–	–	–
Dsouza et al. (2019)	2019	MNB ^c and TF	–	–	–	–	–	–	–	–	–
Failed (2018)	2018	MaxEnt ^d and n-gram	–	–	–	–	✓	–	–	–	–
Rani and Kumar (2017)	2017	Lexical-based	–	–	–	–	–	–	–	–	–
Aung and Myo (2017)	2017	Lexical-based	–	–	–	–	–	–	–	–	✓
Nasim et al. (2017)	2017	RF, ^e TF-IDF, and lexical-based	–	–	–	–	–	–	–	–	✓
Ullah (2016)	2016	SVM and n-gram	–	–	–	–	✓	–	–	–	–
Dhanalakshmi et al. (2016)	2016	NB and TF-IDF	–	–	–	–	–	–	–	–	–
Welch and Mihalcea (2016)	2016	SVM and lexical-based	–	–	–	–	–	–	–	–	–
Altrabshch et al. (2014a)	2014	SVM and n-gram	–	–	–	–	✓	–	–	–	–
Altrabshch et al. (2014b)	2014	SVM and unigrams	–	–	–	–	–	–	–	–	–
El-Halees (2011)	2011	SVM and TF-IDF	–	–	–	–	–	–	–	–	–

Table 1 (continued)

System [Ref]	Year	Approach	Word embedding	Word sense disambiguation	Sentiment-encoded word embedding	Sentence types	Sequential processing	Sentiment dictionary combination	Sentence types	Contextual polarity	
										sentiment shifter rules	Negation handling
Mac Kim and Calvo 2010	2010	NMF ^f (Lee and Seung 1999) and lexical-based	-	-	-	-	-	-	-	-	-
^a Artificial Neural Network											
^b Term Frequency											
^c Multinomial Naïve Bayes											
^d Maximum Entropy											
^e Random Forest											
^f Non-negative Matrix Factorization (NMF)											

based on the traditional machine learning-based approach and feature selection algorithm for classifications. DTLP aims to address the aforementioned issues and solve some of the complex linguistic problems. DTLP has the same objective as the aforementioned related works but differs in terms of method and process representation. To the best of our knowledge, a method that integrates deep learning-based method (*CNN*, *BiLSTM*, and *attention mechanism*), word embedding, sentiment shifter rules, statistical, linguistic, and sentiment knowledge has not been thoroughly studied for solving students' feedback sentiment analysis problem. Furthermore, to our knowledge, in the domain of student feedback, there is little work that goes beyond the sentiment analysis by attempting to also reveal potential reasons when classifying to negative/positive context.

3 Research methodology

We aim to classify students' feedback polarity. To do this, we designed and developed a method called DTLP. The methodology contains the following steps: First, data gathering and data pre-processing of collected data are performed using NLP techniques to remove special characters and stop words form a feature space. Next, we introduce the applied techniques and then present the DTLP. Figure 1 provides a quick summary of the DTLP which presents the process of classification. The following subsections describe each step in detail. Finally, we assess the performance of the DTLP, and the results obtained during the study will be presented and discussed. We also introduce different widely used evaluation metrics to measure the performance of the DTLP. Furthermore, we explore the combination of different variables or features to obtain the best performance. Moreover, we also explore the potential of a deep learning-based approach and the traditional machine learning-based approach for classification.

4 Proposed method

The DTLP system architecture is presented in Fig. 1. It shows the full functionality of the system to classify students' feedback into "*positive*" and "*negative*". The system architecture is composed of four main modules, namely: (i) "input student feedback"; (ii) "pre-processing"; (iii) "feature extraction"; (iv) "student feedback analysis"; and (v) "visual reports". The input student's feedback module takes a text as input. Then, the pre-processing module is applied to the text for further processing. The pre-processed text is sent to the feature extraction module to extract features and create the corresponding feature vector. Finally, as shown in Fig. 1, by analysing the text, the system assigns an

appropriate label to each sentence. In the following sections, we discuss each module of the DTLP system in more depth. Drill down technique is enabled to explore potential causes of classification (e.g. “resource X was not useful” classified into negative context will further visually link to the concepts of potential causes, e.g. “hard to follow” and “irrelevant”).

4.1 Input student feedback module

The first module receives a student’s feedback and then sends it to the pre-processing module. DTLP takes only English language text into account.

4.2 Pre-processing module

At the current step, a student’s feedback is pre-processed using basic linguistic functions to reduce the computational complexity and convert a text into a vector of features that can be analysed through a deep learning-based algorithm. Briefly, pre-processing contains the following tasks: Tokenization (*to split a sentence into several words/tokens*), stop word filtering (*to eliminate words that do not give much information for text analysis (e.g. conjunctions, articles, etc.)*), and stemming (*to obtain the root of each word*). It employs WordNet (Miller and Charles 1991) to get the stem of each word. A WordNet is a dictionary that includes nouns, verbs, adjectives, and adverbs. This dictionary contains 121,962 words. Part-of-speech (POS) tagging (*to label each token in a sentence that can be “verbs”, “adjectives”, etc.*).

4.3 Vector representation module

A feature is an attribute that could be used to detect patterns in large datasets. A vector representation refers to steps to convert an input text into a set of features to specify whether a sentence is a “pos” or “neg” sentence. A vector is specified as $X_t = \{x_1, x_3, x_4, x_5, \dots, x_n\}$ where each element is set according to the corresponding features. However, we employed the following steps to extract the required features.

4.3.1 Sentiment knowledge-based features

The pipeline of sentiment analysis consists of the following steps: (i) sentiment dictionary combination; (ii) applying sentiment shifter and contextual polarity rules; (iii) determining the type of sentence; and (iv) computing sentence sentiment score.

- (i) *Sentiment dictionary combination*: It aims to build a high-coverage sentiment dictionary to increase the number of words in a restricted lexicon (*called*

integration several sentiment dictionaries (ISD)). To do this, it combined several sentiment dictionaries with various formats and sizes (e.g. General Inquirer (Stone and Hunt 1963), AFINN (Nielsen and “A new ANEW 1103), Opinion Lexicon (Hu and Liu 2004), SenticNet4 (Cambria et al. 2016), SO-CAL (Taboada et al. 2011), NRC Hashtag Sentiment Lexicon (Mohammad et al. 1308), Sentiment140 Lexicon (Mohammad et al. 1308), WordNet-Affect (Strapparava and Valitutti 2004), SentiWordNet (Baccianella et al. 2010), and Subjectivity Lexicon (Riloff and Wiebe 2003)). These lexicons were developed manually or mechanically to categorize positive and negative ideas in a document. We explained in detail the processes to combine various sentiment dictionaries in our previous papers ((Abdi et al. 2020, 2018)). Meanwhile, we also use the Semantic Sentiment Technique (SST) to expand the sentiment dictionary coverage.

Semantic Sentiment Technique (SST) as we explained, word coverage is the main limitation of a sentiment dictionary. To tackle this problem, we employ the SST approach to allow calculating the emotional word score that does not exist in ISD. To do this, given a word, firstly, we collect all synonymous words using WordNet, $WS = \{W1, W2, W_n\}$ where W_n is a synonymous word. Secondly, SST accomplishes the following steps for each word W_n of WS using a loop: (i) *If the word exists in ISD and the sentiment score is positive*, add (+1) to (Pes_{sw}) and (ii) *if the word exists in ISD and the sentiment score is negative*, add (−1) to (Nes_{sw}). Finally, Eq. (1) calculates the total sentiment word score as follows:

$$S_{sw} = \frac{1}{n} \times \sum Pes_{sw} - \frac{1}{m} \times \sum Nes_{sw} \quad (1)$$

where the number of positive and negative words is represented by m and n; accordingly, Pes_{sw} being the positive emotional score, and Nes_{sw} being the negative emotional score.

- (ii) *Sentiment shifter and contextual polarity rules*: A dictionary-based method uses a lexicon to classify an opinion/view into positive or negative opinion. On the other hand, since the polarity of a word relies on the context in which it emerges, the pre-defined polarity of a word in the lexicon can be modified. Thus, to cope with this issue, we consider the following rules in sentiment analysis.

Negation handling: In some cases, specific words (e.g. “not”, “hardly”, “None”, “Seldom”, etc.) change a negative sentence into a positive one or the other way around. As an example,

in sentence “*I do not like this course*”, the negation word, “*do not*” will change the polarity of the word “*Like*” (*positive to negative*). Thus, the polarity of the sentence is negative.

Sentiment shifter handling: It also refers to some contrary words (e.g. “*but*”, “*except for*”, “*nevertheless*”, etc.) that reverse the polarity of a sentence following them. In other words, the polarity of a phrase preceding a contrary word and following it is opposite to each other. Thus, to determine the polarity of a sentence, we only take into account the phrase after a contrary word. To achieve this, we employed various strategies proposed by Nguyen and Nguyen (2017) to deal with the aforementioned problems. For instance, given the sentence “*The course was interesting but it was hard to find all study materials*”, the word (“*but*”) will reverse the polarity of the phrase “*The course was interesting*”. The polarity of each phrase can be: “*The course was interesting [+1]*, *but it was hard to find all study materials [-1]*”. However, the polarity of the given sentence is negative.

- (iii) **Sentence types:** Since the sentence type (e.g. “*conditional/interrogative*”, “*subjective/objective*”, and “*sarcastic*”) affects the performance of a method to identify the polarity of a sentence (Chen et al. 2017), we also consider it in our sentiment analysis process. We explain each of them as follows:
 - (a) **An objective and subjective:** An objective sentence does not contain a sentiment word, while a subjective sentence contains an emotional statement/word and represents an idea/opinion (“*It was nice topic*”).
 - (b) **Interrogative and conditional:** A text containing a sentiment word may not express an idea/opinion (Narayanan et al. 2009; Liu 2012). For instance, given the sentences “*Was the module organised well?*” and “*If the course topic is interesting, I will take it*”, both of them consist of a sentiment word, but they do not indicate a positive or negative idea on “*module*” or “*course*”.
 - (c) **Sarcasm:** It is referred to the mock. In other words, sarcasm, unlike negation, uses positive words to express negative polarity. For example, as shown in the sentence “*I would like to join a class on holidays*”, the surface is a positive opinion (“*like*”), but the whole sentence indicates a negative opinion. In other words, there is a conflict between the statement “*on holidays*” and the articulation “*like*”. Sarcasm identification is generally very difficult, as

in comparison with the negation and shifter rules, the lexical features cannot provide enough information to identify it. DTLP employed a set of heuristic rules proposed by Bharti et al. (2015) to identify sarcasm. In order to obtain more information in detail, the reader is referred to the appendix, section “Heuristic rules”.

- (iv) **Sentiment score calculation:** DTLP firstly determines the type of sentence and considers the negation and shifter rules. Secondly, it looks for a word in the ISD to get its sentiment score. Meanwhile, DTLP uses SST to compute the sentiment score of a word, if the word does not exist in ISD. However, the total sentiment score of a sentence is calculated using the sum of the sentiment score of the words.

We consider the following features for each sentence:

- How many words are polarized positively;
- How many words are polarized negatively;
- The total sentiment score of a sentence;
- Frequency of individual negation words;
- If a sentence is subjective;
- If a sentence is objective;
- If a sentence is interrogative;
- If a sentence is conditional;
- Sentiment-encoded word embedding.

4.3.2 Statistical and linguistic knowledge-based features

We extract the essential attributes from a sentence as follows:

Word embeddings feature is a text representation approach that can be used for feature learning. It maps words into a vector space of a small dimension, which is assigned to real value. Word embeddings indicate the meaning of the words. It specifies to which extend the words are semantically identical. Furthermore, it is capable of capturing both syntactic and semantic information of words. In our work, we used word2vec (W2V) representation of text. It represents a word as a vector having 300 dimensions. Word2vec approach proposed by researchers at Google. W2V uses a shallow neural network to process a text. The Word2Vec neural network takes a text document as input and produces a vocabulary with each word having a vector associated with it. Word2vec includes the continuous bag-of-words (CBOW) model (Mikolov et al. 1301) and the skip-gram model (Mikolov et al. 2013). The skip-gram forecasts source context words based on target words,

whereas the CBOW anticipates target words based on origin context words.

Word embeddings drawbacks along with the application of the W2V approach in various NLP tasks and its advantages, the W2V approach also has two major drawbacks. We used the following strategies to cope with the issues.

- (a) *Word sense disambiguation (WSD)*: The W2V method is unable to differentiate between several meanings of words (Kamkarhaghighi and Makrehchi 2017). Hence, it generates only one representation per word form. As an example, given two sentences “the teacher watches the play” and “the teacher went out to play”, the word “play” has multiple meanings in each phrase according to context (e.g. “play: verb”; “play: noun”). To cope with the aforementioned issue, we employed the POS tagger as a feature for each word. To achieve this, we concatenate a six-dimensional binary vector (“preposition”, “conjunction”, “noun”, “verb”, “adjective”, and “adverb”) with the associated W2V vector.
- (b) *Sentiment-encoded word embedding*: The W2V also does not take into account the sentiment knowledge of a word. This means that given two sentences (e.g. “He is a weak presenter” and “He is a strong presenter”), the word embedding approach maps two words (“weak” and “strong”) into close vectors in the embedding space, while both words appear in similar contexts and have opposite polarity. However, in terms of sentiment analysis, both words have different vector representations. Therefore, to solve the current issue, we combine sentiment knowledge and the W2V process. To do this, a binary vector of word polarity (“positive” and “negative”) is attached to the corresponding word embedding.

We extract the following features:

- Word embedding features include d-dimensional vector embedding, six-dimensional binary vector, and two-dimensional binary vector.
- Part-of-speech (POS) tagging: frequency of nouns, adjectives, verbs, and adverbs.
- Punctuation’s feature: frequency of exclamation (“!”) and question (“?”) mark.

However, the input features are classified into two main groups as follows:

- (a) Word-level feature.

- Word embedding features include d-dimensional vector embedding, six-dimensional binary vector, and two-dimensional binary vector.

- (b) Features of sentence level.

- How many words are polarized positively;
- How many words are polarized negatively;
- The total sentiment score of a sentence;
- Frequency of individual negation words;
- Part-of-speech (POS) tagging: frequency of nouns, adjectives, verbs, and adverb;
- Punctuation’s feature: frequency of exclamation (“!”) and question (“?”) mark;
- If a sentence is subjective;
- If a sentence is objective;
- If a sentence is interrogative;
- If a sentence is conditional.

4.4 Student’s feedback analysis module

Sentence classification layer assigns each sentence a class label. Therefore, it generates a collection of labelled sentences. As shown in Fig. 1, the current layer uses a non-linear model which exploits various features and long-range dependencies across multiple steps to classify each student’s feedback. In this module, we used the CNN layer (*including convolutional, pooling, and flatten layers*), BiLSTM layer, attention layer, concatenation layer, and fully connected layer. The CNN layer is initialized by the word-level features. After convolutional and pooling process the input, the output is fed into the BiLSTM layer. BiLSTM processes the received features and generates a hidden state at each step ($h_{i0 \leq i \leq n}$, n : the number of hidden states or the number of outputs). However, for more processes, the BiLSTM is followed by the attention mechanism layer as follows:

Attention mechanism layer: In a sentence, different words can have different contributions. Therefore, we leverage the attention mechanism to learn the importance of words in a sentence and combine the representations of these words to compose a sentence vector. Furthermore, the attention mechanism assigns different attention to different words during the learning of sentence representations. In the other words, the aim of the attention mechanism is to let the method know which part of the input sentence is important during the training so that the method pays attention to the most informative information and effective features (Bahdanau et al. 1409). The attention layer is defined as follows:

$$e_i = \tanh(W.h_i + b) \quad (2)$$

$$\alpha_i = \text{Soft max}(W'.e_i + b') \quad (3)$$

$$v_i = \sum_{i=1}^n (\alpha_i.h_i) \quad (4)$$

where W , W' , b , and b' are parameter vectors that are learnt during the training. The SoftMax (.) and tanh (.) are activation functions that are used to determine the output of each node in an artificial neural network. Equation (2) is a traditional one-layer feedforward neural network where the tanh activation function transforms each state of BiLSTM to obtain e_i as a new hidden representation of h_i . Furthermore, the SoftMax function is applied to get attention weights α_i as shown in Eq. (3). Finally, the new sentence vector v_i is computed using Eq. (4).

The concatenation layer integrates the extracted vector representation (v_i) in attention layer and sentence-level features to form a final vector representation. Finally, the current vector is passed into the fully connected layer to reveal the final classification or predict a corresponding label.

A brief description of the applied techniques: The reader is referred to the appendix of this manuscript in which we explain briefly the underlying mechanisms of RNN, LSTM, bidirectional LSTM, CNN, Naïve Bayes (NB), decision tree (DT), artificial intelligence network (ANN), support vector machine (SVM), and random forest (RF).

4.5 Feedback summarization results module

To simplify the analysis of student comments about different issues such as course, teacher performance, classroom, library, and project, DTLP includes a visualization stage to generate word clouds, opinion classification, opinion summarization, as well as graphs of changes in students' feedback over time. This stage is an important part of DTLP since users can use it for the teaching and learning improvement process cycle; as a response to feedback a teacher may change or improve the course content or material; it can, in addition, be helpful to curriculum developers and administrators.

As shown in Fig. 1, students' comments are used as inputs by DTLP, which produces feedback analysis visualizations as outcomes that could be beneficial for decision making. For reporting, the column graph (Fig. 2) presents the percentage of positive and negative student feedback. This stage also presents the classification of students' comments (Table 2) in terms of "Positive" and "Negative". In addition, DTLP would be able to create a summary of positive and negative students' feedback (Table 3) separately. It aims to extract formative information from student feedback. Meantime, summarization is the procedure of

condensing a huge number of student comments into a concise form. A word cloud (Fig. 3 (1)) is a visual representation of text data or words. It is used to highlight popular words in different sizes. The size and colour of a word show how important it is. The word cloud can be used as a search reference by the users. For instance, if a teacher wants to know what comments have been mentioned for the word "Lecture" (which is highlighted in green colour), the teacher can enter a search entry, e.g. "Lecture" to obtain the corresponding students' feedback. Moreover, a word cloud can help to identify whether students are taking an interest regarding a special issue such as course and teacher. In other words, it shows that students used the most positive words or negative words while giving feedback. Furthermore, given a word (e.g. "Hard"), the DTLP also identifies the number of students' feedback sentences that include the words "Hard", its alternative synonyms and other words associated with it with high frequency (e.g. "Study", "Lecture", "course", "Laboratory", "course material", "teacher knowledge", etc.). It is worth noting that the number will be represented on the relationship line between two words (Fig. 5 (2)). Exploiting these relationships will not only allow informing a teacher what aspects/elements were perceived as challenging for students but also about their potential sources. Finally, as shown in Fig. 5 (3), the corresponding sentences related to the selected words are presented. For instance, the corresponding sentences of words "hard" and "Lectures" are shown in Fig. 5 (3) to facilitate the interpretation of the system output through visual representation, which makes the output of our system relevant for a learning analytics dashboard to support teachers.

Figures 4 and 5 show the most frequently used words (e.g. unigram and bigram words) in positive and negative students' feedback. As shown in Fig. 4, positive feedbacks include crucial words such as "Good", "Excellent", "Fine", "Good, library", and "Nice". Similarly, Fig. 5 shows that negative feedbacks are reflected by words such as "Bad", "Poor", "Hard", "Problem", and "Insufficient, book", which obtained a high-frequency score. However, these words indicate which terms or words are responsible to classify positive and negative feedbacks.

5 Experimental results and discussion

In the following subsections, we introduce the results of the experiment that we performed to verify the effectiveness of the presented method (DTLP). The purpose of the experimental assessment is to answer the following questions: (i) *Can DTLP solve the problem of large-scale student feedback classification?* and (ii) *in comparison with the existing*

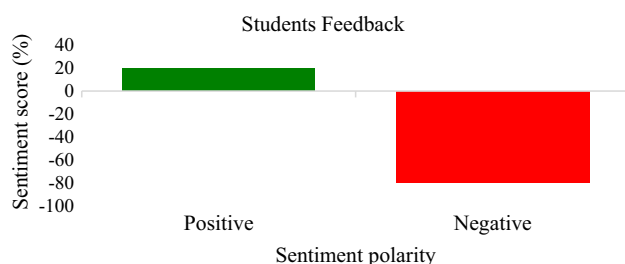


Fig. 2 Sample of percentage of positive and negative students' feedback sentences

methods, can DTLP improve the performance of sentiment classification in various datasets?

For this purpose, we carried out different experiments as follows:

- (1) Since our method uses different resource information (e.g. word embedding, word-level feature, and sentence-level feature), we study the effect of different feature sets. Hence, the performance of DTLP with a combination of different feature sets is evaluated.
- (2) The performance of the DTLP is compared with various traditional methods based on machine learning approaches based on deep learning (with variants of a baseline LSTM model and a CNN model). Meantime, we also evaluated the performance of DTLP with and without the attention mechanism layer.
- (3) The performance of the DTLP is also compared to those of newly released and well-known studies.

We first present the dataset used for the experiment. Then, the measures used to assess the performance of the DTLP are described. Furthermore, an explanation of each experiment and the way how they are chained during the experiments are presented. Finally, we present a discussion of these results and the corresponding implications.

5.1 Dataset

We used a comprehensive education dataset (students' feedback) in our experiment. The following student comments datasets: Welch and Mihalcea (Welch and Mihalcea 2016), online survey¹ (*Students' feedback collected from student surveys administered by a university in the Netherlands*), SASF² and Van Nguyen, et al. (Failed 2018) as displayed in Table 4. In order to evaluate the performance of the presented method, we needed a gold standard data, which would consist of all right outcomes. To achieve this,

¹ <https://docs.google.com/forms/d/1C8MvuPBF131ii1drkg3T90okD4K6rx8me3FB43GewI8/>.

² <https://www.kaggle.com/chandusrujan/sentimental-analysis-on-student-feedback>.

the annotators used the opinionated sentences to create the gold standard with the following process: splitting the text into sentences and then tagging the opinion of each sentence with polarities "Pos" (*to express satisfaction*) or "Neg" (*to express dissatisfaction*). Before applying any method for classification tasks and performing any experiment on the dataset, applying pre-processing steps is necessary to prepare the dataset. Table 5 presents some samples of comments and the corresponding label.

5.2 Performance measurement

We used four evaluation metrics such as accuracy, recall (R), precision (P), and F1-measure to evaluate the performance of the proposed method (Manning et al. 2008). These metrics are calculated using the following four indicators: true-positive (TP) and true-negative (TN) values. It means that both experts and the DTLP classifier assigned positive and negative labels, respectively, to sentences. False positive (FP) means that the sentences had negative labels but were incorrectly categorized as positive, and similarly, false negative (FN) means that the sentences had positive labels but were incorrectly categorized as negative. However, using these definitions, we calculate the performance measures using the following equations:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{F1-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (8)$$

5.3 Classifier method

The DTLP (as presented in Fig. 1) uses CNN, BiLSTM, and attention layer for sentence classification. We use the aforementioned dataset for the training. Firstly, each sentence is represented using the word embedding, sentiment, linguistic, and statistical knowledge features to convert an input to the feature set. Then, the CNN layer receives the word-level feature vector and creates an output. The CNN layer outcome is sent to the BiLSTM layer. A feature vector extracted from the BiLSTM is passed to the attention layer. Finally, the integration of the extracted vector representation (v_i) in attention layer and sentence-level features is passed into a fully connected layer to classify the sentence sentiment orientation. An activation function of the transfer function adjusts the non-linearities of the hidden layers (e.g. "ReLU"). Additionally, it is a simple function that returns

Table 2 Sample of students' feedback sentences classification

Negative students' feedback	
Student's comment	It started very not organized and improved a bit as the module went on, was definitely annoying in the beginning
Student's comment	It was unclear where to find what information and where lectures would take place, this also improved as the module went on, in my opinion some kind of document was needed of where to find what and when to do what and what to read for when, this was missing i think
Student's comment	It was organized very badly
Student's comment	Some teachers refused to record lectures despite them giving required information during these lectures while being hard to hear
Student's comment	This module was super unstructured
Student's comment	It was hard to find all your study materials, the lectures were given in different places
Student's comment	Math wise it seems like we need a ton of previous knowledge on proofs
Positive students' feedback	
Student's comment	Yes, project groups were nice and the company integration too, to improve the process
Student's comment	Guest lectures-nice visual, interesting topic, makes your brain think in the business and it space

Table 3 Sample of student's feedback summarization

Do You Want to Create a Summary of Students Feedback (Y/N): y	
<i>[Summary of Negative Students Feedback]</i>	
It was unclear where to find what information and where lectures would take place, this also improved as the mc went on, in my opinion some kind of document was needed of where to find what and when to do what and what read for when, this was missing I think. It was hard to find all your study materials, the lectures were given in different places	

the node's output. This network is trained by an optimization algorithm (e.g. "adam") with loss functions (cross-entropy (negative log-likelihood)). We have also used a learning rate (i.e. $lr=0.03$) to minimize the loss function. To create a robust system, we also use the l2 constraints regularization and dropout approach (Hinton et al. xxxx). Dropout is a neural network training technique which is used to overcome the overfitting problem. In other words, a dropout is a training approach in which randomly chosen neurons are disregarded (Kim xxxx). The dropout rate is a meta-parameter that must be defined by the user. For regularization, we additionally apply a l2 norm restriction during training (Zhang and Wallace 1510). L^2 norm³ is computed by taking the square root of the sum of the square vector values: $\|v\|_2 = \sqrt{a_1^2 + a_2^2 + \dots + a_n^2}$, where $\vec{v} = (a_1, a_2, \dots, a_n)$.

Hyperparameter Tuning: Hyperparameter optimization has a key role in deep learning where we need to set several parameters, such as activation functions, kernel size, filters, pooling method, strides size, dropout regularization, learning rate, momentum, batch size, epochs, optimization algorithms, weight initialization, and the number of neurons in the hidden layer. To adjust the hyperparameters of our trained model, we employ the random search capability. Random search is a technique to assess a model for various parameter combinations. In our model, we used the Python programming language and its libraries such as "Numpy", "sklearn", and "Math". The deep learning is implemented through the Keras which runs on top of the TensorFlow. The main emphasis was to examine different parameters in order to obtain the best results. Random search was employed for the following objectives:

³ <https://machinelearningmastery.com/vector-norms-machine-learning/>.

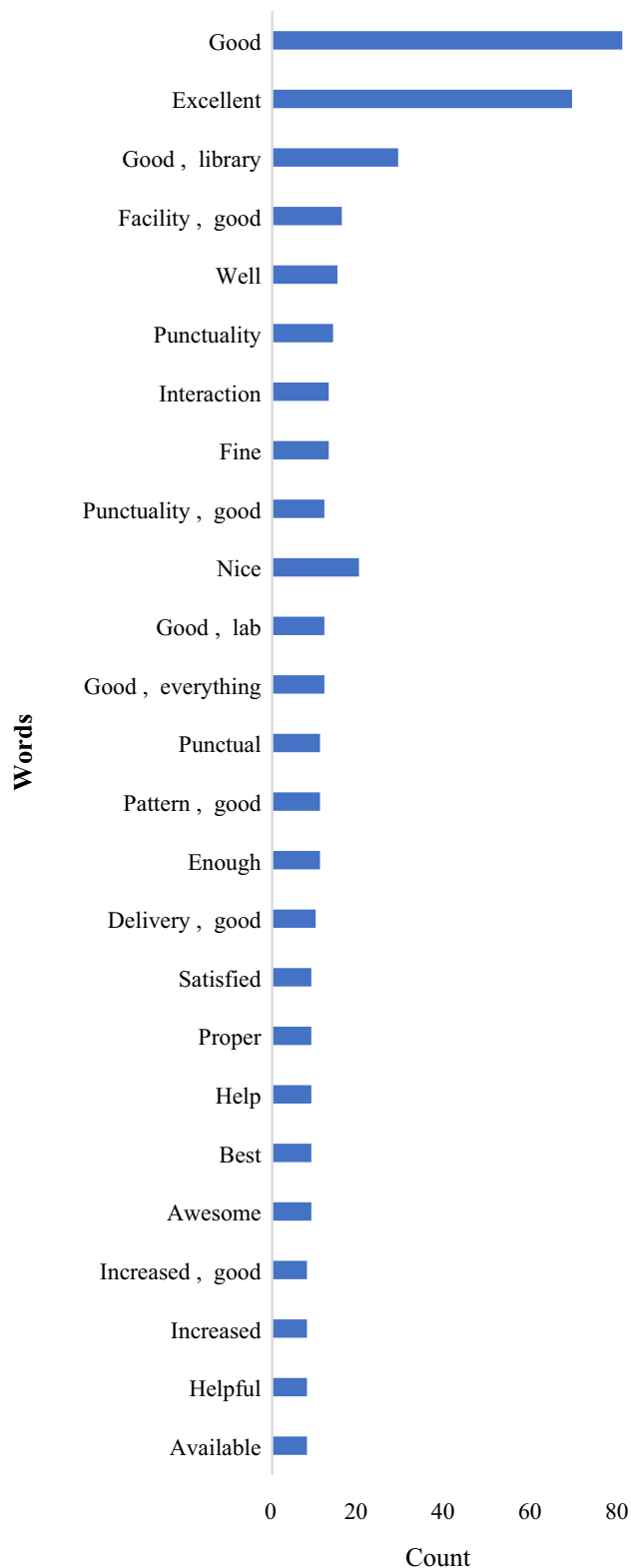


Fig. 4 Most frequently used words in positive students' feedback

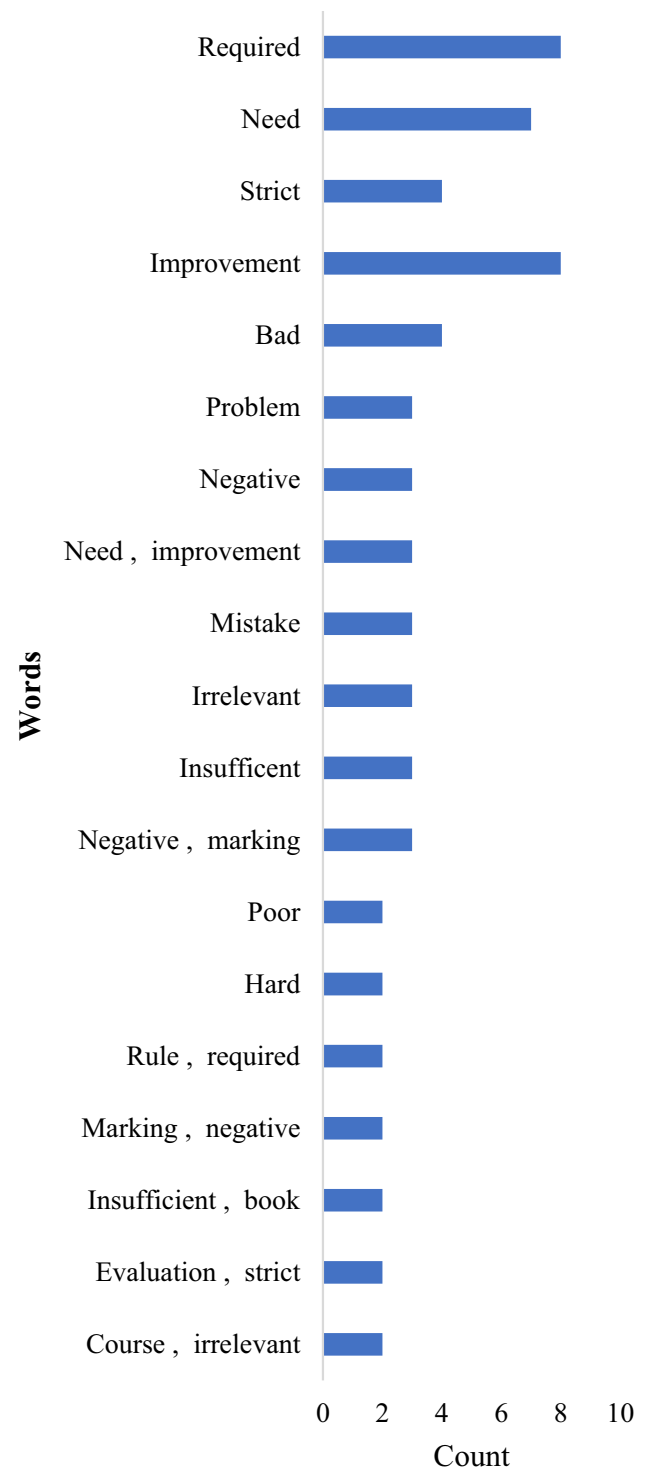


Fig. 5 Most frequently used words in negative students' feedback

5.5 Performance comparison with Naive models and deep learning-based methods

Performance comparison of several supervised algorithms such as traditional machine learning-based method, deep learning-based method, and DTLP is shown in this

Table 4 An overview of the dataset that was utilized

Dataset	Number of students feedback sentences
Welch and Mihalcea (Welch and Mihalcea 2016)	1042
Van Nguyen, et al. (Failed 2018)	186
SASF	16,000
Survey	300

Table 5 Samples of student feedbacks in dataset (“1”=Pos, “-1”=Neg)

Target	Student feedback
1	Excellent lectures are delivered by teachers and all teachers are very punctual
-1	Good and qualified teachers but few lacks in interaction with students and fails to describe a topic
1	Lectures help in clear our concept
-1	Faculties are doing quite well but need more senior lecturers
1	Teachers are enthusiastic to help students during their practice

section. For this purpose, various classification methods were employed as follows:

LSTM: Classification based on the LSTM (the model captures sequential information from one side). First, LSTM is initialized by the vector representation extracted from the dataset, then the sentence classification (SC) is performed through a fully connected layer (FCL).

BiLSTM: Compared to LSTM, BiLSTM is capable of capturing information on both sides (see appendix, Fig. 2). First, BiLSTM is initialized by the vector representation extracted from the dataset, then the SC is done through a FCL.

CNN: First, CNN is initialized by the vector representation extracted from the dataset, then the SC is conducted through an FCL.

CNN+LSTM: Sentence classification based on the CNN and LSTM. First, CNN is initialized by the vector representation extracted from the dataset, then the LSTM is located on top of the CNN layer. Finally, the SC is carried out through an FCL.

CNN+BiLSTM: Unlike the previous method (CNN+LSTM), the BiLSTM is located on top of the CNN layer. Finally, the SC is performed through an FCL.

BERT (Devlin et al. 1810): We utilize a pre-trained bidirectional encoder representation from transformers (BERT) model using an online released TensorFlow library.⁴

CNN+BiLSTM+Attention layer: As an enhancement of the CNN+BiLSTM model, this model adds an attention mechanism to assign different weights to feature extracted from BiLSTM for better representing importance. However, TC is done based on the CNN, attention mechanism, and

BiLSTM. First, CNN is initialized by the vector representation extracted from the dataset, then the BiLSTM based on attention mechanism is located on top of the CNN layer. Finally, SC is doing through the fully connected network.

Finally, the sentence classification is performing based on the traditional machine learning-based methods (SVM, RF, DT, and NB) and artificial neural networks (ANN) (a neural network with multiple hidden layers).

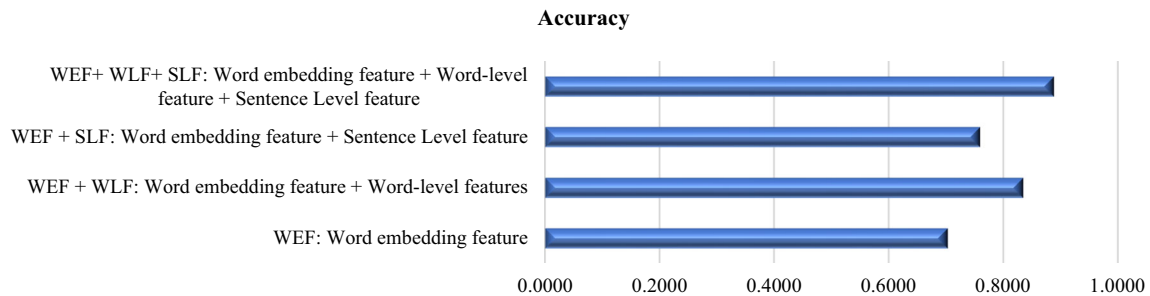
The effectiveness of the aforementioned methods is compared using F-measure, precision, recall, and accuracy. The comparison results of DTLP and other methods on the dataset are shown in Table 7. According to the results demonstrated in table, the SVM performs best (63.33%/accuracy value) among the other traditional machine learning methods (e.g. ANN, RF, DT, and NB). In comparison with the SVM algorithm, DTLP improved its performance as follows: 37.00% (F-measure) and 40.17% (accuracy). However, as an overall comparison, we conclude that in terms of accuracy and F-measures, the DTLP consistently outperforms the baselines methods. It can be due to some reasons such as unlike a traditional machine learning-based method, a deep learning-based method can manage larger datasets more effectively. Additionally, a deep learning-based method uses several hidden layers for training; hence, the corresponding model can learn more and more. Furthermore, a traditional machine learning-based method does not consider the semantic relationships between words in a sentence, while a deep learning-based method (e.g. BiLSTM) considers the word order or syntactic information of a sentence.

Furthermore, as shown in Table 7, the model DTLP obtains the best performance. In other words, the DTLP significantly outperforms the other deep learning-based methods and displays greater performance with high

⁴ <https://github.com/tensorflow/tensor2tensor>.

Table 6 The results obtained from the experiments carried out varying the DTLP method

Methods	WEF	WLF	SLF	Precision	Recall	F1-measure	Accuracy
DTLP <i>FULL</i>	+	+	+	0.9074	0.8909	0.8991	0.8878
DTLP <i>WEF + WLF</i>	+	+		0.8776	0.8600	0.8687	0.8333
DTLP <i>WEF + SLF</i>	+		+	0.8182	0.8000	0.8090	0.7571
DTLP <i>WEF</i>	+			0.7442	0.7619	0.7529	0.7000

**Fig. 6** The results achieved from the experiments carried out varying the feature set

accuracy as compared with other models. The DTLP obtained the best result ($F\text{-measure}=0.8991$ and $accuracy=0.8878$) in comparison with CNN-BiLSTM, which does not use the attention mechanism layer. Generally, the deep learning-based methods obtained better results in comparison with the Naive models. Moreover, the performance of the DTLP will be improved when it uses the attention layer.

The column (“DTLP improvement”) in Table 7 displays the relative improvement when comparing the DTLP technique to other methods. The following formula calculates the relative improvement: $\frac{(\text{OurMethod} - \text{Othermethod})}{\text{Othermethod}} \times 100$. The symbol “+” indicates that DTLP enhances the corresponding method performance by improving the accuracy. As an example, the DTLP enhances the performance of the CNN-LSTM with 5.26 $F\text{-measure}$ and 7.90 $accuracy$. Furthermore, the DTAM improved the $F\text{-measure}$ and $accuracy$ as follows: $(4.16\% \leq F\text{-measure} \leq 83.27\%)$ and $(5.03\% \leq Accuracy \leq 71.43\%)$, respectively.

We also show a comparison of DTLP with several models using Fig. 7. The results from the table are also aggregated into bar graphs to understand these results at a higher level. Each bar indicates the $F\text{-measure}$ and $accuracy$ of each method, which are RF, SVM, ANN, CNN, etc. From the plot, it can be seen that the models’ performance in view of $F\text{-measure}$ and $accuracy$ varies across the different methods. Figure 7 presents that a bigger bar demonstrates high performance and thus is a better model. However, a deep learning-based method obtains better results in comparison with the traditional machine learning-based methods.

5.6 Comparison with related methods

In this section, the performance of DTLP is compared with other recently published papers, which have been applied to the students’ feedback data to classify students’ comments. Table 8 shows the results of the DTLP model against other methods, mainly measured by F-measure and accuracy. As evident from Table 8, (Misuraca et al. 2021) is the worst method when compared with the other methods. In comparison with (Tamrakar et al. 2102; Katragadda et al. 2020; Failed 2019a) and (Dsouza et al. 2019), the DTLP model significantly outperforms other methods and displays greater robustness with high performance. The DTLP obtained the best result ($F1\text{-measure}=0.8991$ and $accuracy=88.78\%$) in comparison with (Failed 2019a), which performs the best and has $F\text{-measure}$ and $accuracy$ measures of 0.7838 and 71.43%, respectively. In other words, the DTLP can classify the student’s feedback sentences with significantly higher performance.

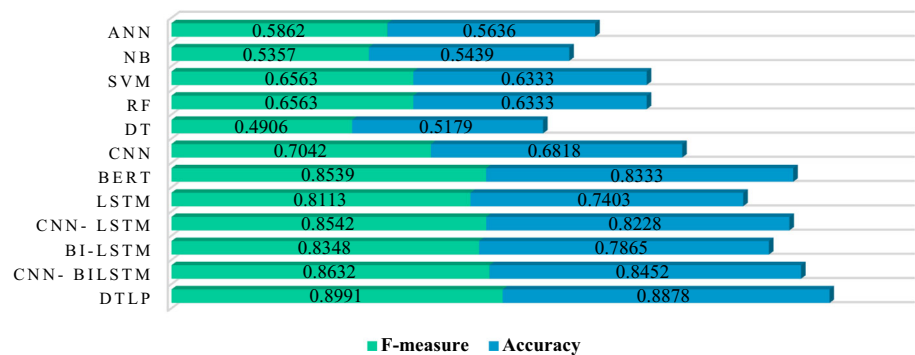
Table 8 presents the relative improvement measure. For example, the DTLP improves the performance of (Failed 2019a) with 14.71% $F\text{-measure}$ and 24.29% $accuracy$. Furthermore, the DTLP improves the performance as follows: $(14.71\% \leq F\text{-measure} \leq 83.90\%)$ and $(24.29\% \leq accuracy \leq 86.01\%)$.

5.7 Discussion

From the experiments above (Tables 6, 7, and 8), we observed that the DTLP outperformed other methods and exhibits higher accuracy. This is because:

Table 7 Different methods applied on dataset

Group	Methods	Precision	Recall	F1-measure	Accuracy	DTLP improvement (%)	
						F-measure	Accuracy
Other	SVM	0.5833	0.7500	0.6563	0.6333	37.00	40.17
	ANN	0.5152	0.6800	0.5862	0.5636	53.37	57.50
	RF	0.5455	0.6667	0.6000	0.5789	49.85	53.34
	DT	0.4643	0.5200	0.4906	0.5179	83.27	71.43
	NB	0.5357	0.5357	0.5357	0.5439	67.83	63.23
Recurrent	BiLSTM	0.8136	0.8571	0.8348	0.7865	7.70	12.87
	LSTM	0.7818	0.8431	0.8113	0.7403	10.82	19.92
	BERT	0.8261	0.8837	0.8539	0.8333	5.29	6.53
CNN	CNN	0.6579	0.7576	0.7042	0.6818	27.67	30.20
	CNN-LSTM	0.8542	0.8542	0.8542	0.8228	5.26	7.90
DTLP	CNN-BiLSTM	0.8542	0.8723	0.8632	0.8452	4.16	5.03
	CNN- BiLSTM+ Attention layer	0.9074	0.8909	0.8991	0.8878	–	–

Fig. 7 The performance comparisons between different models

- (i) The DTLP method combines different resource information that takes advantage of sentiment, word embedding, statistical and linguistic knowledge-based information to improve the student's feedback sentences classification. Table 6 shows that the DTLP model's performance is significantly better on the unified set of features than the one which learns from a subset of features. In other words, the results demonstrate that DTLP can provide valuable insights into the relationships between various resource information.
- (ii) The word embedding approach leads to the extraction of deep semantic relationships between words and augments other features extracted using sentiment, linguistic, and knowledge. The DTLP employed a word embedding approach, while other methods (Table 8, excluding the method presented by Failed 2019a) do not use w2v features.
- (iii) Despite the advantage of using a word embedding approach, there are certain drawbacks with the word2vec method including i) a word embedding

approach based on vector representation does not include enough sentiment-related knowledge to perform sentiment analysis. Hence, words having opposite polarity, such as “like” and “dislike”, are mapped to word vectors that are close together. Therefore, the DTLP combines the sentiment knowledge and W2V approach to tackle the aforementioned problems, while other methods do not take into account these current disadvantages of the W2V approach. For this purpose, i) the DTLP combines several sentiment dictionaries (*as prior lexical knowledge*) as means of overcoming the limitation of word coverage and ii) the word embedding approach also is not able to distinguish the senses of a word. Unlike other methods (*refer to Table 8*), DTLP can handle this issue.

- (iv) DTLP considers the contextual polarity (*e.g. negation*), shifter rules (*e.g. but clause*), and types of sentences (*e.g. sarcastic sentence*), while other methods disregard these rules (Table 8, excluding

Table 8 The performance of the DTLP against other relevant works in the literature

Methods	Precision	Recall	F1-measure	Accuracy	DTLP improvement (%)	
					F-measure	Accuracy
DTLP	0.9074	0.8909	0.8991	0.8878	—	—
Tamrakar, et al. (Tamrakar et al. 2102)	0.5556	0.6667	0.6061	0.5738	48.35	54.72
S, et al. (Failed 2019a)	0.7632	0.8056	0.7838	0.7143	14.71	24.29
Dsouza, et al. (Dsouza et al. 2019)	0.5000	0.6400	0.5614	0.5283	60.15	68.04
Katragadda, et al. (Katragadda et al. 2020)	0.5000	0.6800	0.5763	0.5455	56.02	62.76
Misuraca, et al. (Misuraca et al. 2021)	0.5238	0.4583	0.4889	0.4773	83.90	86.01

the method presented by Misuraca et al. (2021) and Katragadda et al. (2020) (*negation handling*).

- (v) Most of the existing proposed methods use a BOW approach for a vector representation. However, a BOW-based approach does not consider the word order and semantic relationships between words. In other words, the BOW-based approach could not distinguish between a positive and negative sentence such as “A course was difficult” and “No course was difficult”. A BOW approach also suffers from high sparsity (features values as zero) and high dimensionality. Meanwhile, some proposed algorithms used the n-gram approach (Lin and ”Rouge 2004) for vector representation. Although the n-gram considers the word order in comparison with the BOW approach, the n-gram works well in a long text, unlike the short text, since long text has a large number of co-occurring words. To tackle the problems, DTLP employs the word embedding method and BiLSTM method for sequential data processing (*e.g. word order in a sentence*).
- (vi) According to the results in Table 7, we observed that the traditional machine learning (ML)-based method obtained the worst results. This can be due to several reasons such as feature engineering (*a process to extract important features from a dataset*) that must be done by an expert. The performance of a traditional ML-based method depends on how accurately the features are selected and extracted. However, unlike the traditional ML-based methods, a deep learning-based method can automatically extract features and learn from the data. Additionally, the main difference between deep learning and machine learning-based methods is data dependency. It means that an ML-based method is appropriate for a small amount of data, while a deep learning-based method needs a large amount of data to understand it perfectly and

perform well. Furthermore, a machine learning method has its limitation when it comes to capturing dependence between words in a sentence because it assumes that each word is independent of the other. Therefore, the expectation is that a recurrent neural network is better suited for the DTLP. Technically, it can handle word order, consider long-range dependencies and guarantee better performance than other methods. The attention layer helps DTLP to focus on a certain part of the given information, highlight effective features, and assign different weights to all outputs of BiLSTM to find accurately the sentiment orientation of a given input sentence.

5.8 Limitations

The major limitations related to this work include the pre-trained word embedding method, which is a google pre-trained model that contains public online data. Therefore, it consists of information about domains other than education and teaching information, which may not be adequate for our specific aim.

5.9 Implications

In this work, our hypotheses were (i) a combined feature set constructed using feature vectors based on word embedding, sentiment knowledge, linguistic and statistical knowledge can potentially improve the performance of classification, and (ii) a joint network with CNN, BiLSTM, and attention layer is capable to generate a significant performance. To prove the hypotheses, the potential of employing a deep learning-based method was explored. To do this, the DTLP, a method based on the CNN, BiLSTM, and attention mechanism, has been proposed and applied to the corresponding dataset. We observed that the integration of several valuable resource information had a promising impact on classification. The experiments also showed very

promising results. Our findings showed that the proposed model could be considered to classify students' feedback sentences. This proposed method extracts positive and negative opinions and relevant information to provide various information about related issues including fine-grained sentiments linked to various concepts/potential causes. It can be used to explore how to improve the teaching and learning process. This work proves useful in verifying learning processes in terms of various perspectives such as teaching, course content, facilities, resources, and assessment. The results can help to prepare useful information for teachers, students, and educational institutions that can be used in daily practices. Limitations of this work can serve as suggestions to researchers to take the new challenges in future work.

5.10 Implications for the practical use

The proposed system can be used as an assistant tool to enhance the quality of education in several areas such as teaching material and adjusting teaching approaches during learning process. More specifically, teachers can reflect on the received feedback and take the required activities to enhance instruction and educational programmes. By getting indication on students' perceptions during learning in the form of concepts reflected in their feedback and their relationships, e.g. a specific learning resource or a procedure linked to a specific sentiment and their causes, as well as having the opportunity to detect specific groups of learners that may benefit from targeted feedback, a teacher can react earlier during a learning process, compared to situations when teachers get the chance to analyse feedback after the course has ended. Furthermore, with the introduced method, teachers can benefit from making informed decisions about their learning process adjustment needs specifically for larger groups of students which can be challenging otherwise considering limited teacher resources and the amount of manual work and time they needed to spend on processing student feedback. One major benefit of our system is that the combination of the techniques it exploits allows to not only detect the fine-grained sentiments related to different aspects of a teaching/learning process but also drill down to potential causes.

6 Conclusion and future work

The advancement in technologies has innovated the education field dramatically. Deep learning-based approaches have attracted lots of attention recently in many fields including education. In this work, DTLP was proposed to classify students' comments into positive and negative sentence. DTLP takes advantage of the CNN, BiLSTM, and

attention mechanism, where coarse-grained local features are produced by CNN, BiLSTM takes into account the sequential processing, attention mechanism highlights discriminative and effective features.

DTLP contains the following main parts: (i) input vectors, (ii) CNN layer, (iii) BiLSTM layer, (iv) attention layer, and (v) final vector representation and a fully connected layer. Firstly, several NLP techniques are applied to the students' feedback sentences as a pre-processing step. Then after that, DTLP generates features (e.g. word embedding, sentiment knowledge, linguistic and statistical knowledge) from unstructured text to create an input vector. Secondly, the CNN layers are initialized by the word-level feature. After convolutional and pooling processes, the CNN layer outcomes are sent to the BiLSTM. Furthermore, the attention layer produces a new sentence vector representation. To produce a final vector representation, the retrieved vector representation from the attention layer and sentence-level feature is combined using the concatenation layer. Finally, the current vector is sent to the FCL to reveal the final classification or predict a corresponding label.

We have conducted experiments on the datasets to validate our method. First, we compare different DTLP methods as follows: $DTLP_{WEF+SLF+WLF}$, $DTLP_{WEF+SLF}$, $DTLP_{WEF+WLF}$, and $DTLP_{WEF}$. The findings present that between all of the different DTLP techniques, the $DTLP_{WEF+SLF+WLF}$ obtains the best performance in comparison. Furthermore, the result indicates that those features are complementary to each other. Additionally, the performance of the DTLP was also tested and compared with various supervised-based methods. The obtained results validate our method, and we find that the DTLP outperforms other baseline methods. Moreover, the DTLP with attention layer obtained significant performance in comparison with the DTLP without attention layer.

In summary we also compared DTLP with existing proposed methods. The results demonstrated that the DTLP performed the best across all methods. In particular, our results suggest that the model that learnt from the unified feature set could achieve higher performance than one that learnt from a feature subset.

In future work, a new feature set as an additional independent variable to improve the classification model will be considered. Another avenue for future work: to investigate and develop a method using various deep learning-based methods. Furthermore, we would like to consider "student assessment" (Casalino et al. 2021) as a new component of our system. Moreover, we aim to apply idea presented in paper (Camastra et al. 2020) to the "Feedback Summarization Results module" to enhance graph-based knowledge representation.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00500-023-07926-2>.

Authors' contributions All authors contributed to the study conception and design. AA and GS helped in conceptualization, methodology, investigation, software, writing—original draft, data curation, formal analysis, writing—reviewing and editing, and validation. SM, vdB worked in writing—reviewing and editing. JvH and BV contributed to funding acquisition and project administration. All authors read and approved the final manuscript.

Funding This work is conducted at the IEBIS and CODE research centres and was supported by the BMS faculty grant of the University of Twente. The authors would also like to thank the anonymous reviewers who have contributed enormously to this work.

Data availability The data will be available if the user request.

Declarations

Conflict of interest I hereby and on behalf of the co-authors, declare all the authors agreed to submit the article exclusively to this journal and also declare that there is no conflict of interests regarding the publication of this article.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References Abdi A, Hasan S, Shamsuddin SM, Idris N, Piran J (2020) A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion. *Knowl-Based Syst* 213:106658

Abdi A, Shamsuddin SM, Hasan S, Piran J (2018) Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment. *Expert Syst Appl* 109:66–85

Altrabsheh N, Cocea M, and Fallahkhair S “Sentiment analysis: towards a tool for analysing real-time students feedback,” In *2014a IEEE 26th international conference on tools with artificial intelligence*, 2014a: IEEE, pp. 419–423.

Altrabsheh N, Cocea M and Fallahkhair S “Learning sentiment from students’ feedback for real-time interventions in classrooms,” In: *International conference on adaptive and intelligent systems*, 2014b: Springer, pp. 40–49

Araque O, Corcuera-Platas I, Sánchez-Rada JF, Iglesias CA (2017) Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Syst Appl* 77:236–246

Aung KZ and Myo NN “Sentiment analysis of students’ comment using lexicon based approach,” In: *2017 IEEE/ACIS 16th*

International conference on computer and information science (ICIS), 2017: IEEE, pp. 149–154

Baccianella S, Esuli A, Sebastiani F (2010) Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *Lrec* 2010(10):2200–2204

Bahdanau D, Cho K, and Bengio Y “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014

Bharti SK, Babu KS, and Jena SK “Parsing-based sarcasm sentiment recognition in twitter data,” In: *2015 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, 2015: IEEE, pp. 1373–1380

Camastra F, Ciaramella A, Maratea A, Son LH and Staiano A “Semantic maps for knowledge management of web and social information,” In: *Computational intelligence for semantic knowledge management*: Springer, 2020, pp. 39–51

Cambria E, Poria S, Bajpai R, and Schuller B, “SenticNet 4: a semantic resource for sentiment analysis based on conceptual primitives,” In: *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers*, 2016, pp. 2666–2677

Casalino G, Castellano G, Vessio G (2021) Exploiting time in adaptive learning from educational data. *International workshop on higher education learning methodologies and technologies online*. Springer, pp 3–16

Chauhan GS, Agrawal P, and Meena YK “Aspect-based sentiment analysis of students’ feedback to improve teaching–learning process,” In *information and communication technology for intelligent systems*: Springer, 2019b, pp. 259–266

Chen T, Xu R, He Y, Wang X (2017) Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN. *Expert Syst Appl* 72:221–230

Derick L, Sedrakyan G, Munoz-Merino PJ, Kloos CD, Verbert K (2017) Evaluating emotion visualizations using affect vis, an affect-aware dashboard for students. *J Res Innov Teach Learn* 10 (2):107–125

Devlin J, Chang M-W, Lee K and Toutanova K “Bert: pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018

Dhanalakshmi V, Bino D, and Saravanan AM, “Opinion mining from student feedback data using supervised learning algorithms,” in *2016 3rd MEC international conference on big data and smart city (ICBDSC)*, 2016: IEEE, pp. 1–5.

Dsouza DD, Deepika DPN, Machado EJ, Adesh N (2019) Sentimental analysis of student feedback using machine learning techniques. *Int J Recent Technol Eng* 8(14):986–991

El-Halees A “Mining opinions in user-generated contents to improve course evaluation,” In: *International conference on software engineering and computer systems*, 2011: Springer, pp. 107–115

Hinton GE, Srivastava N, Krizhevsky A, Sutskever I and Salakhutdinov RR “Improving neural networks by preventing co-adaptation of feature detectors,” *arXiv e-prints*. [Online]. Available: <https://ui.adsabs.harvard.edu/abs/2012arXiv1207.0580H>

Hu M and Liu B “Mining and summarizing customer reviews,” In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177

Kaewyong P, Sukprasert A, Salim N, and Phang FA “The possibility of students’ comments automatic interpret using lexicon based sentiment analysis to teacher evaluation,” In: *3rd International conference on artificial intelligence and computer science (AICS2015)*, 2015, pp. 179–189

Kamkarhaghghi M, Makrehchi M (2017) Content tree word embedding for document representation. *Expert Syst Appl* 90:241–249

Katragadda S, Ravi, V, Kumar P and Lakshmi GJ “Performance analysis on student feedback using machine learning algorithms,” In: *2020 6th international conference on advanced computing*

- and communication systems (ICACCS), 2020: IEEE, pp. 1161–1163
- Kim Y “Convolutional neural networks for sentence classification,” *arXiv e-prints*. [Online]. Available: <https://ui.adsabs.harvard.edu/abs/2014arXiv1408.5882K>
- Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791
- Lin C-Y “Rouge: A package for automatic evaluation of summaries,” In: *Text summarization branches out*, 2004, pp. 74–81
- Liu B (2012) Sentiment analysis and opinion mining. *Synthesis Lectures Human Lang Technol* 5(1):1–167
- Lwin HH, Oo S, Ye KZ, Lin KK, Aung WP, and Ko PP “Feedback analysis in outcome base education using machine learning,” In: *2020 17th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON)*, 2020: IEEE, pp. 767–770
- Mac Kim S and Calvo RA “Sentiment analysis in student experiences of learning,” In: *EDM*, 2010: Citeseer, pp. 111–120
- Manning CD, Raghavan P, and Schütze H “Introduction to information retrieval? Cambridge university press 2008,” *Ch*, vol. 20, pp. 405–416
- Mikolov T, Chen K, Corrado G and Dean J “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- Mikolov T, Sutskever I, Chen K, Corrado GS and Dean J “Distributed representations of words and phrases and their compositionality,” In: *advances in neural information processing systems* 26, Burges CJC, Bottou L, Welling M, Ghahramani Z, and Weinberger KQ Eds.: Curran Associates, Inc., 2013, pp. 3111–3119
- Miller GA, Charles WG (1991) Contextual correlates of semantic similarity. *Lang Cognit Process* 6(1):1–28
- Misuraca M, Scepi G, Spano M (2021) Using Opinion Mining as an educational analytic: an integrated strategy for the analysis of students’ feedback. *Stud Educ Eval* 68:100979
- Mohammad SM, Kiritchenko S, and Zhu X, “NRC-Canada: building the state-of-the-art in sentiment analysis of tweets,” *arXiv preprint arXiv:1308.6242*, 2013
- Narayanan R, Liu B and Choudhary A “Sentiment analysis of conditional sentences,” In: *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 180–189.
- Nasim Z, Rajput Q and Haider S “Sentiment analysis of student feedback using machine learning and lexicon based approaches,” In: *2017 international conference on research and innovation in information systems (ICRIIS)*, 2017: IEEE, pp. 1–6
- Nguyen H and Nguyen M-L “A deep neural architecture for sentence-level sentiment classification in twitter social networking,” In: *International conference of the Pacific association for computational linguistics*, 2017: Springer, pp. 15–27
- Nielsen FÅ “A new ANEW: evaluation of a word list for sentiment analysis in microblogs,” *arXiv preprint arXiv:1103.2903*, 2011
- Rani S, Kumar P (2017) A sentiment analysis system to improve teaching and learning. *Computer* 50(5):36–43
- Riloff E and Wiebe J “Learning extraction patterns for subjective expressions,” In: *Proceedings of the 2003 conference on Empirical methods in natural language processing*, 2003, pp. 105–112
- S. S, D. T, and D. B, “Analysis of Student Feedback using Deep Learning,” *International Journal of Computer Applications Technology and Research*, vol. 8, pp. 161–164, 05/07 2019a, doi: <https://doi.org/10.7753/IJCATR0805.1004>.
- Sedrakyan G “Process-oriented feedback perspectives based on feedback-enabled simulation and learning process data analytics,” 2016.
- Stone PJ and Hunt EB “A computer approach to content analysis: studies using the general inquirer system,” In: *Proceedings of the May 21–23, 1963, spring joint computer conference*, 1963, pp. 241–256.
- Strapparava C and Valitutti A “Wordnet affect: an affective extension of wordnet,” In: *Lrec*, 2004, vol. 4, no. 1083–1086: Citeseer, p. 40
- Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. *Comput Linguist* 37(2):267–307
- Tamrakar L, Shrivastava D and Ghosh DS “Student sentiment analysis using classification with feature extraction techniques,” *arXiv preprint arXiv:2102.05439*, 2021
- Ullah MA “Sentiment analysis of students feedback: a study towards optimal tools,” In *2016 International workshop on computational intelligence (IWCI)*, 2016: IEEE, pp. 175–180
- Van Nguyen K, Nguyen VD, Nguyen PX, Truong TT, and Nguyen NL-T “UIT-VSFC: Vietnamese students’ feedback corpus for sentiment analysis,” In *2018 10th International conference on knowledge and systems engineering (KSE)*, 2018: IEEE, pp. 19–24.
- Welch C and Mihalcea R “Targeted sentiment to understand student comments,” In: *Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers*, 2016, pp. 2471–2481
- Zhang Y and Wallace B “A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification,” *arXiv e-prints*, p. [arXiv:1510.03820](https://ui.adsabs.harvard.edu/abs/2015arXiv151003820Z). [Online]. Available: <https://ui.adsabs.harvard.edu/abs/2015arXiv151003820Z>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.