Exploring semantic similarity in English learners' texts through topic modelling

Natalia Grabar[1,†], Olena Yurchenko[2,*,†], Olga Cherednichenko[3,†] and Arsenii Lukashevskyi[4,†]

[1] CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
[2] CNRS, Maison Européenne des Sciences de l'Homme et de la Société (MESHS, UAR3185), 365 bis, rue Jules Guesde 59650 Villeneuve d'Ascq, France
[3] Bratislava University of Economics and Management, Furdekova 16, Bratislava, Slovak Republic
[4] National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv, 61002, Ukraine

Abstract

Assessing semantic similarity between texts of different length, such as questions and their extended answers, remains challenging in natural language processing. This study investigates whether topic modelling, specifically BERTopic, can effectively capture semantic similarity in such cases. The EFCAMDAT corpus, a large-scale dataset of learners' written texts, was used for experimentation. The research addresses two key questions: (1) Do students' texts correlate with the questions they are asked? (2) How does this correlation vary across different levels of foreign language proficiency? The findings indicate semantic similarity between questions and answers can be identified using topic modelling. Keyword analysis confirms a correlation between the examined elements; however, the method for determining semantic similarity still requires further refinement to improve accuracy.

Keywords

text semantic analysis, topic analysis, topic modelling, semantic similarity, text summarisation

1. Introduction

Text semantics evaluation is a strategic area of linguistics for both linguistic theory and natural language processing (NLP). It belongs to AI tasks related to understanding texts in different languages, on the one hand, and, on the other hand, to solving the problem of the lack of processed data in low-resource languages and domains by developing Model Transfer Learning.

Semantic analysis, or the process of recognizing the semantics of a text and establishing relationships between words, phrases, sentences, paragraphs and texts by their language-independent meanings, is an important component in various NLP tasks, such as sense recognition [1], text summarization [2], short text evaluation [3], determining the degree of semantic similarity between texts [4], text classification [5], clustering of text documents [6, 7], Cross-lingual Transfer [8, 9], etc.

The research proposes to focus on assessing the semantic similarity of texts through topic modelling. It is the first step in the search for a formalised unit of meaning that unites texts of different sizes (in our case, questions for English as a Foreign Language (EFL) learners and answers provided by students with varying language levels from the EFCAMDAT corpus) and a tool that allows us to evaluate the results of condensing meaning into a single phrase or even a word, and vice

CLW-2025: Computational Linguistics Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine
* Corresponding author.
† These authors contributed equally.
natalia.grabar@univ-lille.fr (N. Grabar); olena.yurchenko.etu@univ-lille.fr (O. Yurchenko); olga.cherednichenko@vsemba.sk (O. Cherednichenko); arsenii.lukashevskyi@sgt.khpi.edu.ua (A. Lukashevskyi).
0000-0002-0237-4554 (N. Grabar); 0000-0002-6074-0241 (O. Yurchenko); 0000-0002-9391-5220 (O. Cherednichenko); 0009-0001-8178-717X (A. Lukashevskyi)

versa, expanding an idea into a whole text while retaining its main idea. This can be the basis of the Transfer Learning Model for low-resource languages and domains, in particular for Ukrainian.

We hypothesise that meaning transfer can occur between languages, within a language, its stylistic layers, or between common language and professional domains. Thus, if 100 people answer the same question, their answers can be reduced to the essence of the question, i.e. a single element of meaning. The research aims to test how to determine that given texts refer to one question since the answers to this question must have something in common - the elements of meaning we are looking for.

To solve this problem, we use the Topic Modelling method of the modern large language model BERT. As a research material, we use the EFCAMDAT corpus of English texts, which contains answers to questions from learners of English as a second language (L2). This study compares texts of different lengths to one topic. The questions in the EFCAMDAT corpus are examples of short texts, and we call them text-questions or simple questions. Student answers are examples of long texts, and we call them text-answers or answers. We expect the topical modelling method to show us the correlation between the corresponding text-questions and text-answers. We assume that questions are topics and answers are texts that correspond to these topics. Hence, the BERTopic model should match answers and questions with some minimal error. In this way, we want to answer the research questions:

1) Do students' texts correlate with the questions they are asked?

2) How does this correlation occur at different levels of foreign language proficiency?

2. Related Work

The main objective of semantic similarity is to measure the distance between the semantic meanings of a pair of words, phrases, sentences, or documents [4]. Semantic similarity is a metric defined over a set of documents or terms, where the idea of distance between items is based on the likeness of their meaning or semantic content as opposed to lexicographical similarity [10]. We used it to estimate the strength of the semantic relationship between units of language (words), concepts and texts, through Topic Modelling obtained according to the comparison of information from the texts, which are answers to the same question. But it is noted that the term semantic similarity usually includes only the ‘is a’ relationship. In that case, we can also use the notion of 'semantic relatedness' [11]. Defining semantic relatedness also requires understanding the lexical hierarchy [11], using the concept of the lexical-semantic field [12, 13] and methods of measuring it [14]. Lexical semantics is also related to concepts such as connotation (semiotics) [15] and collocation, a specific combination of words that can be or often surround a single word.

In addition, words in the language are combined non-compositionally to form multi-word expressions (syntagms), whose meaning cannot be derived from the standard representation of their components [16]. Thus, for many domains or languages, it is also essential to have not only cross-linguistic representations of individual words but also to compose them into correct embeddings in phrases, sentences, and other high-level cross-linguistic representations [17, 18].

The method of Topic Modelling of texts is relevant today and has great potential for improving the transfer learning model for low-resource languages and domains. However, topic modelling methods such as latent Dirichlet allocation (LDA), latent semantic analysis (LSA), or keyword extraction technique KeyBERT are not suitable because they are based on the “bag of words” principle. Keywords alone are insufficient to solve the task of correlating text with a topic (question). It is necessary to account for the size of the context so that the weights are proportional [19]. BERTopic is better indicated because it's a model considered the most contextualized due to TF-IDF methodology [20, 21].

Semantic analysis also involves identifying features specific to certain linguistic (professional domains) and cultural contexts as far as possible. We see opportunities for semantic analysis of these features through cross-linguistic comparison of the scope of lexical and semantic fields of individual concepts. We consider this a novel approach to the task of cross-lingual transfer.

3. Materials and Methods

3.1. Materials

The material for the study is the EFCAMDAT [30] corpus [22], which consists of students' answers on questions asked. The students come from all over the world and study English as a foreign language (L2). They may have different CEFR levels (A1, A2, B1, B2, C1, C2). The corpus was first released in July 2013.

The EFCAMDAT corpus is an open-access corpus of student work submitted to Englishtown, an online school from EF Education First [31]. The entire Englishtown course offers 16 levels of language proficiency according to common standards such as TOEFL, IELTS and the Common European Framework of Reference for Languages (CEFR) [23].

The EFCAMDAT corpus consists of scripted writing tasks on a specific question at the end of each lesson. The EFCAMDAT corpus does not have direct information on L1 proficiency, so nationality is the closest proxy for L1 proficiency. EFCAMDAT contains data on students of 198 nationalities. The answers of Ukrainian students make up less than 1%, namely 0.11% of the total number of answers.

For this study, we used the data collected for the second release in September 2017, which contains 1,180,310 scripts (with 7,126,752 sentences and 83,543,480 tokens) written by 174,743 students. This text corpus includes information on learner errors, parts of speech, and grammatical relationships. All tasks were evaluated by English teachers. Currently, EFCAMDAT contains teacher feedback for 66% of the answers [24].

As materials, the questions and a certain number of students' answers to these questions were taken from the EFCAMDAT corpus. We assume that the questions are topics and the answers are texts corresponding to these topics. At each level, we have 24 questions (except for C2, where the number of questions is 8).

3.2. Topic modelling methods

First, we use the LDA method to compare the student answers' Topic Modelling to the manual thematic distribution of questions. The LDA method is applied to a BOW representation: information on the frequency of words is exploited, but their contexts are lost [25, 26].

To match students' text-answers to the studied text-questions, we used the Topic Modelling methods of the Modern BERT large language model. ModernBERT is a modernized bidirectional encoder-only Transformer model (BERT-style) pre-trained on 2 trillion tokens of English and data with a native context length of up to 8,192 tokens. ModernBERT's native long context length makes it ideal for tasks that require processing long documents, such as retrieval, classification, and semantic search within large corpora. The model was trained on a large corpus of text and code, making it suitable for a wide range of downstream tasks, including code retrieval and hybrid (text + code) semantic search [21].

For topic extraction, we used BERTopic method, which is based on the TF-IDF statistical matrix and takes better account of contexts than LDA and other topic modelling models [27, 28]. BERTopic [32] generates document embeddings with pre-trained transformer-based language models, clusters these embeddings, and generates topic representations with the class-based TF-IDF procedure. The semantic properties of Text embedding representations allow the meaning of texts to be encoded in such a way that similar texts are close in vector space.

BERTopic generates topic representations in three steps. First, each document is converted to its embedding representation using a pre-trained language model. Then, before clustering these embeddings, the dimensionality of the resulting embeddings is reduced to optimise the clustering process. Finally, from the clusters of documents, topic representations are extracted using a custom class-based variation of TF-IDF [20].

To find the best clusterisation model for EFCAMDAT corpus, we first tested BERTopic methods on the 20NewsGroups [33] dataset, a classic example of topic modelling, one of the three datasets

used to validate BERTopic. 20NewsGroups contains 18,846 news items from English-language forums. This dataset was pre-processed using Galileo [34] by removing punctuation, lemmatisation, and stop words, as well as removing documents containing less than 5 words and empty messages, which amounted to 1,163 messages. The number of analysed articles after cleaning was 17,734 articles in 20 thematic categories.

Our results for Topic modelling of this benchmark 20NewsGroups align with those reported in [27, 28]. In addition, BERTopic demonstrated good coherence and accuracy in formulating topics using 4 keywords compared to topic titles in 20NewsGroups. Although the BERTopic model does not currently have the best results in terms of CV (topic coherence) and TD (topic diversity) metrics, according to [29], it is among the three most accurate and fastest models.

It should be noted that, when testing BERTopic on the benchmark 20NewsGroups, the first topic is "Topic -1", which groups texts that have not been assigned to any of the topics, the so-called *outliers*. That's why we set the distribution to 20 topics (with and without outliers) and 21 topics (with and without outliers) to compare the results with the distribution marked in the benchmark. We also gave the texts the option to be divided into an automatically determined number of topics and obtained the following results:

1) The division into 20 topics with outliers (Topic -1) refers to emissions of almost 40%-50%, and sometimes 60% of articles from each topic.
2) The division of 21 topics, i.e. 20 topics without outliers (Topic -1), assigns different texts to the relevant topics with high confidence.
3) Many proposed topics (Labels) of the 20NewsGroups benchmark coincide with one Topic, which is appropriate for close topics related, for example, to computers (5 labels) or sports (2 labels). In our results, combining space technology with cars and motorcycles or atheism with religion and politics is possible. However, in the distribution of 20 Topics with outliers, the topics were more clearly identified by the Top 4 keywords than those of 21 Topics without outliers. Although distributed with outlier, all other topics have a more accurate characterization by 4 keywords corresponding to the specified topic than when distributed into 21 topics and without outliers.
4) In addition, the 20NewsGroups benchmark was analysed with the BERTopic model, using the previous generation MiniLM vectorizer compared to vectorizer of ModernBERT. The results were similar to the previous ones, with about a third of the documents for each label belonging to Topic -1 (outlier) divided into 20 topics. However, when divided into 21 topics, the documents were more accurately grouped into separate Topics.

Thus, we concluded that the best distribution for Topic analysis of the EFCAMDAT corpus using BERTopic should be 25 topics at A1-C1 levels and 9 topics at C2 levels, so that the topics outside of the Topic -1 outliers correspond to the distribution of the data to questions in terms of number.

4. Experiment

This section describes the results of experiments with EFCAMDAT corpus in the second release of 2017. Firstly, we clean and preprocess the texts. Then, we compare LDA-based topics with manually chosen ones. Finally, we run BERTopic model to evaluate semantic similarity between text-answers and text-questions.

4.1. Preparing corpus EFCAMDAT for the study

For levels B1-C2, the study was conducted on the entire data set. Since the number of answers at the lower levels A1-A2 was too large, the topic analysis was conducted on a random subset evenly selected for each question. In Figure 1, you can see an example of the stratification validation of the proportional random selection of the answers' number for level A1:

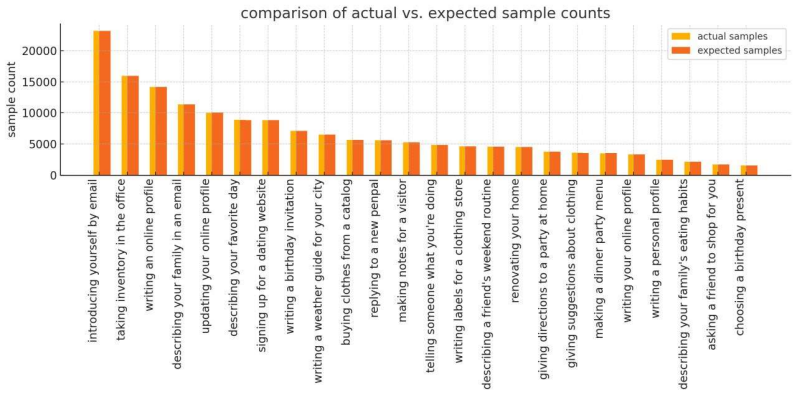


Figure 1: Proportional random selection of the number of answers to questions within level A1.

At the B1 level, some broken lines were removed, which allowed to obtain a more accurate Topic Modelling result.

In data preparation, the texts were comprehensively cleaned: removal of empty answers, lower-case conversion, deletion of non-Latin characters, removal of stop words, including auxiliary verbs. We also cut off answers with poor scores from teachers, i.e. those that received a score below 64 on a 100-point scale. The number of answers remaining after the cleaning is presented in Table 1.

Table 1

Quantitative results of EFCAMDAT corpus cleaning

Level	Nb of questions	Nb answers	Nb answers after cleaning	Rejected answers, %
C2	8	1,940	1,843	5%
C1	24	14,689	14,248	3%
B2	24	61,301	59,358	3%
B1	24	168,138	163,066	3%
A2	24	307,603	163,101	47%
A1	24	625,255	163,205	74%
Total	**128**	**1,178,976**	**564,821**	**52%**

4.2. Topic modelling of EFCAMDAT by LDA

The LDA model was created using Gensim [35], a Python library for topic modelling, where we set the number of topics depending on how many topics the questions were manually divided into at each level of the EFCAMDAT corpus. LDA is trained directly on our EFCAMDAT data.

All the answers in the EFCAMDAT corpus are grouped into subsets that correspond to 128 questions at a particular level of proficiency. There are no common questions across levels. We have also manually identified 10 topics based on questions and indicated at which levels they are, as you can see in Table 2.

Table 2

Manually identified topics, under which 128 EFCAMDAT questions are grouped

№	Themes / Levels	A1	A2	B1	B2	C1	C2
1	Family	+	+	+		+	
2	Business		+	+	+	+	+
3	Travel	+	+	+		+	+
4	Goods	+	+	+	+		

5	Feelings	+		+	+	+	+
6	Habits	+	+	+	+	+	+
7	Learning / Language Training		+	+	+		+
8	Party	+	+	+	+		
9	Home	+			+	+	
10	Health		+	+	+	+	
	Total correspondences:	**7**	**8**	**9**	**8**	**7**	**5**

Using the LDA method, we analysed only levels B1-B2 to see if the approach would be effective for the entire data, taking into account some system slowdowns and redundancy in levels A1-A2. LDA was run in 20 passes with preliminary lemmatisation. We obtained the top 10 words for each topic, a visualisation of the top 30 words, and metrics that can be seen in Table 3:

Table 3

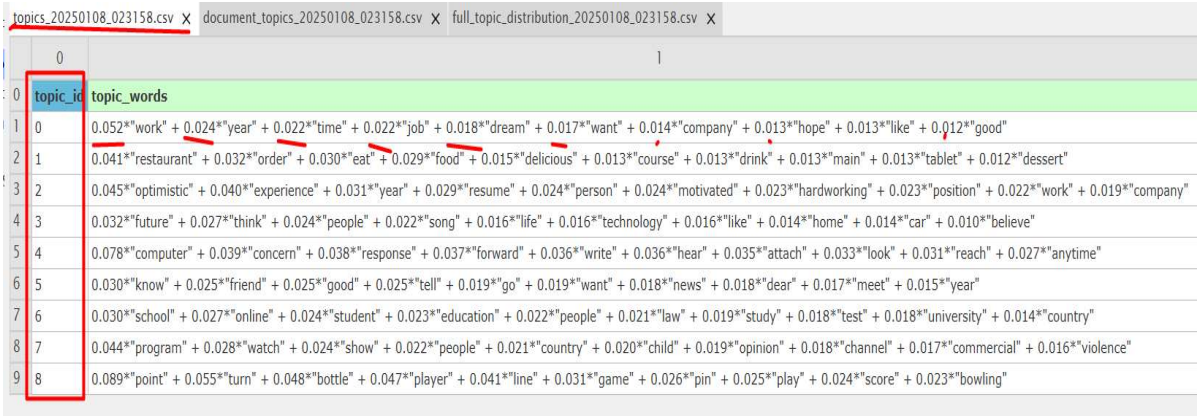
LDA model metrics

Metrics	B1	B2
Coherence Score	0.67	0.60
Nb of answers	168293	61311
Vocabulary size	45841	27820
Average answer length	39.57	56.06

According to the comparison results, the word-topic coherence index is almost the same for both levels and is around 0.6: B1 = 0.66; B2 = 0.59. However, the number of answers at level B1 is 2.74 times higher than at level B2. Therefore, the size of the vocabulary is 1.64 times larger at level B1 than at level B2. At the same time, the length of an answer is 1.4 times longer at B2.

Based on the results of the Topic analysis, 10 keywords were also selected using the LDA method for each of the 9 topics at the B1 level and 8 topics at the B2 level. Their weight indicates how specific and relevant a word is to a given topic compared to other topics.

The keywords were ranked according to their weight and importance within the topic, as can be seen in Figures 1 and 2. A comparison of keywords for the same topics for B1 and B2 shows that at B2, the top 10 words are better when matching manually selected topics.


topics_20250108_023158.csv | document_topics_20250108_023158.csv | full_topic_distribution_20250108_023158.csv

	0	1
0	topic_id	topic_words
1	0	0.052*"work" + 0.024*"year" + 0.022*"time" + 0.022*"job" + 0.018*"dream" + 0.017*"want" + 0.014*"company" + 0.013*"hope" + 0.013*"like" + 0.012*"good"
2	1	0.041*"restaurant" + 0.032*"order" + 0.030*"eat" + 0.029*"food" + 0.015*"delicious" + 0.013*"course" + 0.013*"drink" + 0.013*"main" + 0.013*"tablet" + 0.012*"dessert"
3	2	0.045*"optimistic" + 0.040*"experience" + 0.031*"year" + 0.029*"resume" + 0.024*"person" + 0.024*"motivated" + 0.023*"hardworking" + 0.023*"position" + 0.022*"work" + 0.019*"company"
4	3	0.032*"future" + 0.027*"think" + 0.024*"people" + 0.022*"song" + 0.016*"life" + 0.016*"technology" + 0.016*"like" + 0.014*"home" + 0.014*"car" + 0.010*"believe"
5	4	0.078*"computer" + 0.039*"concern" + 0.038*"response" + 0.037*"forward" + 0.036*"write" + 0.036*"hear" + 0.035*"attach" + 0.033*"look" + 0.031*"reach" + 0.027*"anytime"
6	5	0.030*"know" + 0.025*"friend" + 0.025*"good" + 0.025*"tell" + 0.019*"go" + 0.019*"want" + 0.018*"news" + 0.018*"dear" + 0.017*"meet" + 0.015*"year"
7	6	0.030*"school" + 0.027*"online" + 0.024*"student" + 0.023*"education" + 0.022*"people" + 0.021*"law" + 0.019*"study" + 0.018*"test" + 0.018*"university" + 0.014*"country"
8	7	0.044*"program" + 0.028*"watch" + 0.024*"show" + 0.022*"people" + 0.021*"country" + 0.020*"child" + 0.019*"opinion" + 0.018*"channel" + 0.017*"commercial" + 0.016*"violence"
9	8	0.089*"point" + 0.055*"turn" + 0.048*"bottle" + 0.047*"player" + 0.041*"line" + 0.031*"game" + 0.026*"pin" + 0.025*"play" + 0.024*"score" + 0.023*"bowling"

Figure 2: Key words for the 9 topics of the B1 level.

	0	1
0	topic_id	topic_words
1	0	0.061*"job" + 0.022*"pay" + 0.022*"month" + 0.018*"find" + 0.017*"salary" + 0.016*"income" + 0.016*"year" + ...
2	1	0.020*"company" + 0.019*"meeting" + 0.014*"work" + 0.014*"good" + 0.013*"presentation" + 0.010*"ceo" + 0.010*"new" +
3	2	0.024*"new" + 0.024*"roof" + 0.021*"sunset" + 0.020*"house" + 0.018*"need" + 0.015*"apartment" + 0.014*"property" + ...
4	3	0.039*"year" + 0.034*"work" + 0.030*"job" + 0.029*"sale" + 0.022*"company" + 0.017*"university" + 0.017*"study" + ...
5	4	0.089*"woman" + 0.048*"man" + 0.026*"gender" + 0.023*"work" + 0.018*"law" + 0.018*"think" + 0.015*"country" + ...
6	5	0.031*"laptop" + 0.017*"young" + 0.016*"friend" + 0.015*"push" + 0.014*"go" + 0.014*"plane" + 0.013*"scared" + ...
7	6	0.041*"english" + 0.020*"technology" + 0.019*"read" + 0.017*"use" + 0.016*"book" + 0.015*"language" + 0.014*"internet" +
8	7	0.020*"life" + 0.015*"time" + 0.013*"year" + 0.013*"people" + 0.013*"like" + 0.012*"good" + 0.012*"thing" + 0.010*"live" +
9		

Figure 3: Key words for the 8 topics of the B2 level.

The full probability distribution (Figures 2, 3) for all topics for each answer allows to see how much the answer is related to each topic, not just the dominant one.

According to the results of the LDA thematic analysis, the coherence of answers with the topics that were manually selected is higher at the B2 level (68.84% to 91.25%), compared to the B1 level, where the coherence is quite low (starting at 22.96%, exceeding 60% in only 3 cases and reaching a maximum of 75%).

For example, at the B2 level, it is interesting to note the selection of Topic 0, which shows high coherence (90.35%) and, based on the lexical composition of 10 keywords, can be attributed to the topic Business selected manually.

However, according to the graphs in Figure 4, where this topic is represented by circle 3 and the distribution of vocabulary by 30 keywords, we see that this topic is divided into several lexical and semantic fields that do not even overlap.

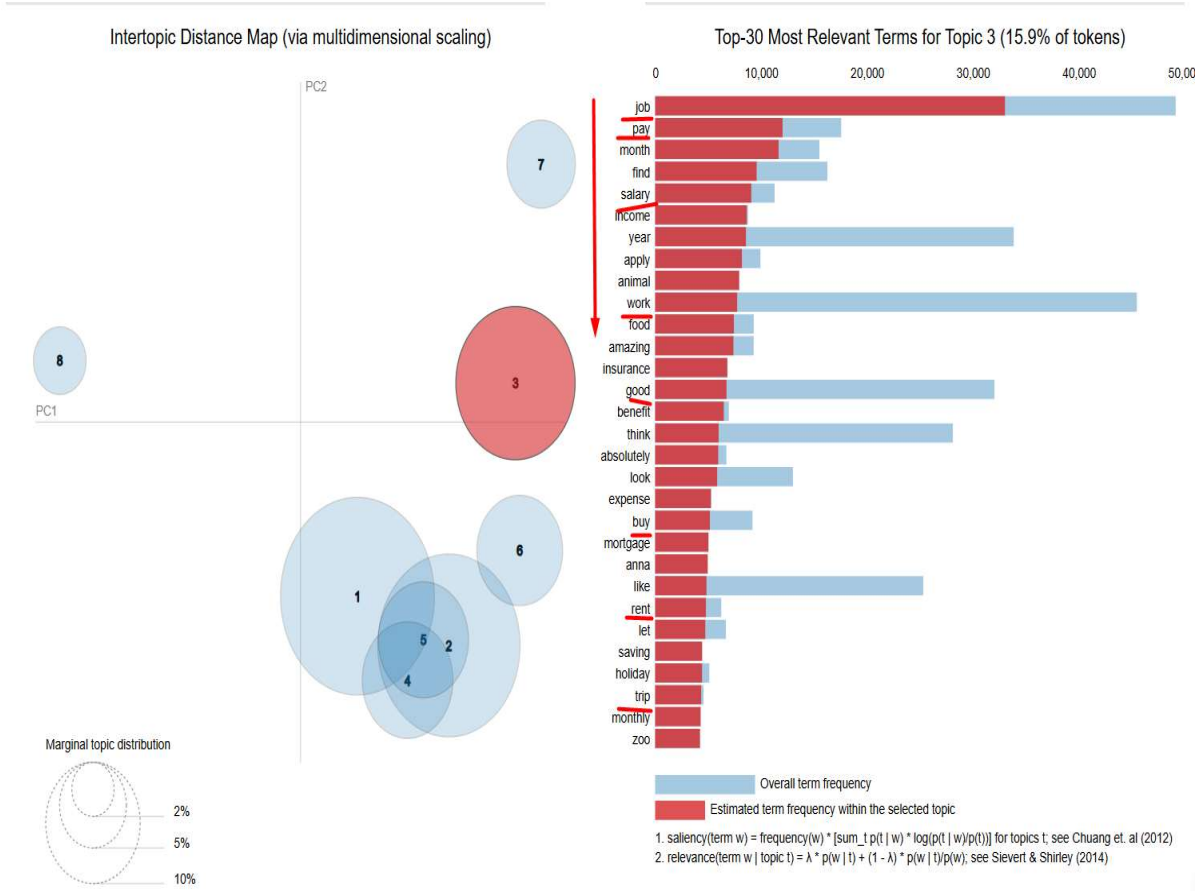


Figure 4: Intertopic Distance Map for level B2.

Thus, we conclude that

1) At level B1, the thematic distribution does not correspond to the 9 manually identified themes per question, in contrast to level B2, where we can correlate some of the themes that were manually identified with the LDA distribution of 8 themes.

2) The visualisations do not always correspond to the given topic numbering id, but they allow us to expand the range of words that are relevant to a topic.

4.3. Topic modelling of EFCAMDAT by BERTopic

Topic modelling was done using Google Colab for all language levels A1-C2 of the EFCAMDAT corpus. We used the ModernBERT-base [36], which has 22 layers and 149 million parameters. The all-MiniLM-L6-v2 was used as a vectorizer for embedding, which provides a good balance between quality and processing speed. To calculate the probability that a certain themes are present in the document, the HDBSCAN model was used at the clustering stage.

4.3.1. Automatic detection of topics in EFCAMDAT with BERTopic

In the first step of Topic Modelling, we did not limit BERTopic to the number of topics found so as not to limit our 'horizons'.

The results of the topic analysis were presented in the form of

- clouds of answers by topic
- hierarchical classification of topics
- a map of distances between topics

As a result, the following number of Topics were identified, which correspond to a certain number of EFCAMDAT questions, as we can see in Table 4:

Table 4

The number of selected Topics in the automatic topic modelling of BERTopic

Level	Nb of Questions	Nb of Topics
A1	24	732
A2	24	506
B1	24	284
B2	24	136
C1	24	83
C2	8	17

The difference in the number of Topics for the same number of Questions by level is explained by the different number of answers, which increases with the lower levels, especially A1-A2.

The resulting visualisations form rather dense areas of topics, which are close to each other, as we can see in Figure 5. Hence, we can assume that these are the semantic fields of the topics set in the themes, which are revealed in the students' answers.

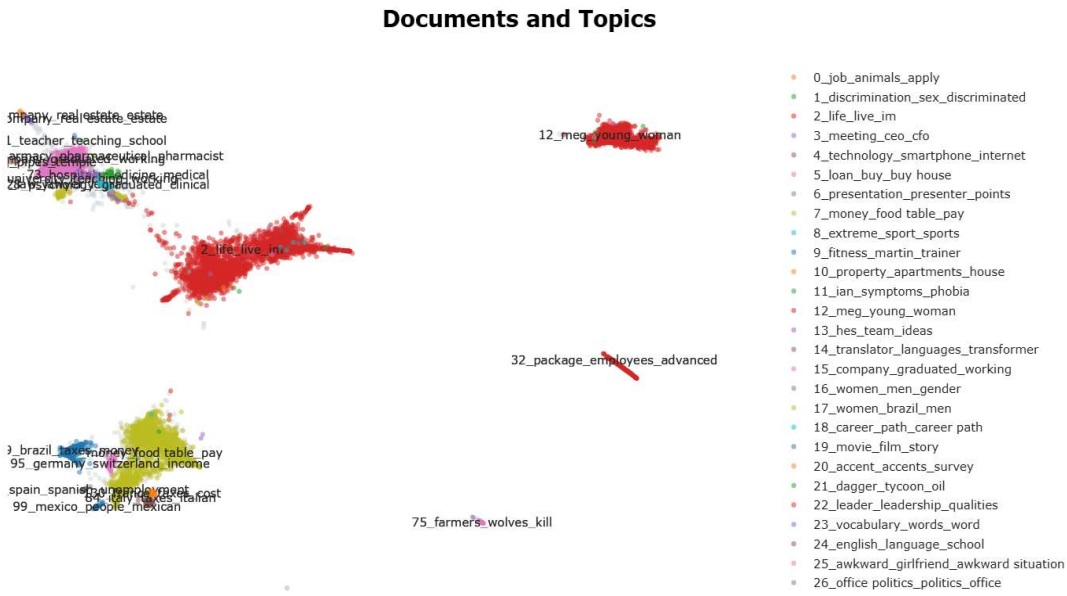


Figure 5: Visualization of the automatic topic distribution of B2 level and a list of topic names presented as Top-3 words.

Comparing the results of the analysis between levels is complicated by the fact that the topics assigned to students at A1-C2 levels do not match. Accordingly, the grouping of topics according to BERTopic modelling does not correspond to the Topics that we identified manually in Table 2.

For example, in Figure 6, red and turquoise groupings at A2 are probably answers to themes: 'A2|Writing a resume', which we assigned to the topic 'Yourself|Family' rather than to the profession, which we did not even select, or 'A2|Complaining about a meal', which was assigned to the general topics 'Food or Feelings', but does not convey an associative connection with the topic of restaurants, which was also not selected in the 10 topics selected manually.

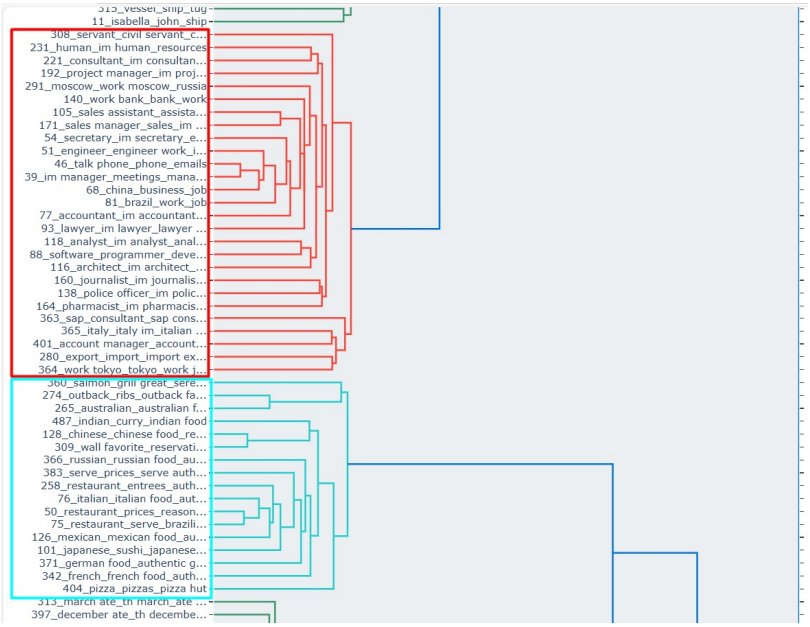


Figure 6: Automatic topic distribution of A2 level.

Since there is no direct correspondence between the manually selected topics and questions, the automatically obtained clusters of topics correlated with the questions can likely lead to another set of topics based on the semes and sememes from lexical and semantic fields.

4.3.2. Identifying 24 and 25 questions in EFCAMDAT by BERTopic

EFCAMDAT answers were divided by BERTopic into 24 topics according to the number of questions at A1-C1 levels. At C2 level, we divide them into 8 topics according to the number of questions.

Then, we correlated the questions of EFCAMDAT with the BERTopics. Considering that BERTopic a priori allocates *Topic -1* for outliers, we also divided answers into 25 topics at A1-C1 levels and 9 topics at level C2 level, with and without outliers.

At this stage, we conduct a quantitative analysis of the Topic Modelling by BERTopic:

- Determine the number of answers for 24 topics identified by Modern BERT and check the quantitative ratio of topics to answers.
- Check whether all answers in the topics are included in the topics, and vice versa;
- Calculate the correlation between the answers and one of the Modern BERT topics.
- Find out if there is an error and what it depends on.
- Compare the analysis data between levels.
- Visualise the results.

In the third final step, we perform a qualitative analysis of the distribution of the maximally cleaned EFCAMDAT:

- Create correlative matrices of Themes and Topics for each level.
- Compare the Topic names defined by 4 keywords from Themes and 10 and 100 keywords from Topics.
- Determine what has a greater impact on the correlation of the text with the given topic.

5. Results

The distribution of answers to EFCAMDAT questions by BERTopic is not homogeneous, as can be seen when comparing the results of the quantitative analysis by levels:

1) **Level C2:** only 2 Topics out of 8 have an accuracy exceeding >98%, 3 Topics >60%, 1 Topic >50%. There are no leaders at this level, as Topic 2 (Topic 0) is >90% and Topic 4 (Topic 2) is >90% for 1 question.
2) **Level C1:** 3 topics out of 24 have an accuracy of >90%, 4 topics >85%, 7 topics >50%. Topic 2 (Topic 0) is in the lead with >90% = 3 questions, >50% = 2 questions.
3) **Level B2:** 6 topics out of 24 have an accuracy of >90%, 7 topics >85%, 9 topics >50%. Topic 2 (Topic 0) is in the lead with >90% = 4 questions, >80% = 2 questions, >50% = 5 questions.
4) **Level B1:** 11 topics out of 24 >90%, 14 topics >85%, 2 topics >82%, 7 topics >50%. Topic 2 (Topic 0) is the leader with >90% = 7 questions, >80% = 3 questions, >50% = 6 questions.
5) **Level A2-A1** data were truncated by 52% and 74%, respectively, so we consider them unsuitable for quantitative analysis.

The distribution of topics by level is not homogeneous: at the lower level B1, there are more questions that correlate with answers by more than 90% compared to level C1.

However, at the same level B1, there are more topics that overlap with each other than at level C1. In Figures 7 and 8, we can see that answers to different questions overlap in Topic 0, while at level C1, there are far fewer such overlaps.

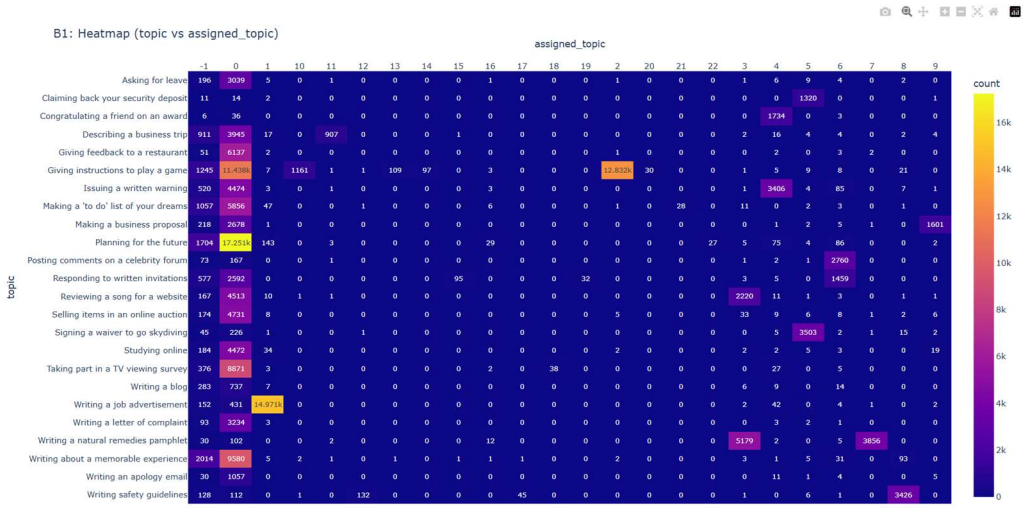


Figure 7: Heatmap B1 - questions EFCAMDAT vs assigned topics BERTopic.

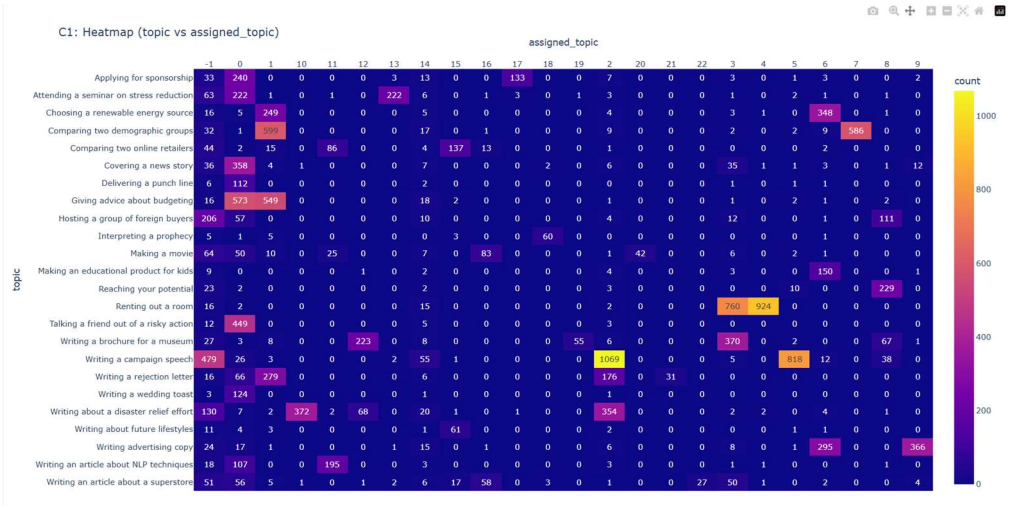


Figure 8: Heatmap C1 - questions EFCAMDAT vs assigned topics BERTopic.

This is also confirmed by the percentage of answers that relate to Themes compared to Topics and is clearly visible in the pie charts (Figure 9).

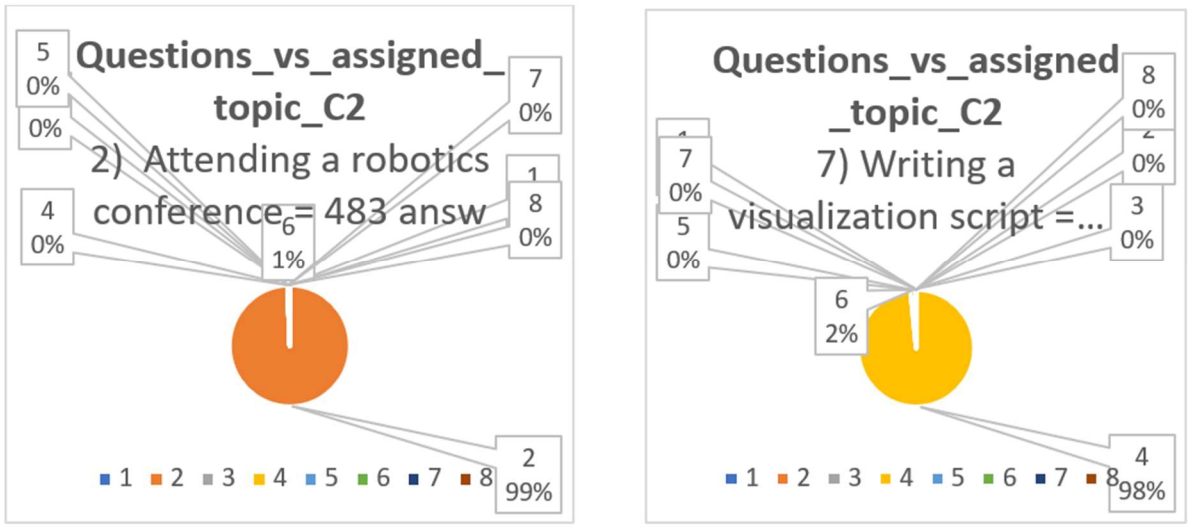


Figure 9: Correlation of BERT topics with EFCAMDAT questions with >90% accuracy.

The quantitative analysis has led to the following conclusions:

1) The lower the level of language learning, the fewer words learners have to express the same idea accurately. The fewer words, the smaller the volume of the lexical and semantic field (LSF) for one concept. That is why words from different lexical and semantic fields are used when revealing Themes in lower levels. Thus, Topics, which are clearly more than 24, can overlap for different tasks, which is also confirmed by the BERTopic distribution.

2) The uneven distribution of Topics may also depend on the wording of the task: more general questions are distributed to a larger number of Topics than more specific ones and vice versa.

3) In most cases, the composition of the top 4 or top 10 keywords in the answers allows us to understand which Theme they relate. This confirms the hypothesis that a question can be reduced to a single concept and it can be represented by a single word, and then expanded to a lexical and semantic field of 4, 10 or even 100 keywords, that make up the concept and can be used in an answer on the same question.

The latter conclusion is also confirmed by the results of the qualitative analysis, which we presented in the form of matrices (Figure 10), where the names of EFCAMDAT Topics are presented horizontally and BERTopics Topics vertically. At their intersection, we see the number of relevant answers that correlate Questions with Topics, and the names of the 4 keywords of the latter can quickly indicate how close this correlation is.

We propose to take the correlation matrix for level C1 in Figure 10 as an example. The same trends emerge from other levels with 24 topics and level C2 with 9 Topics.

It should be noted that there is a small number of outlier responses, which, depending on the type of division into 24 or 25 Topics with outlier or no_outlier, varies between 5-8% depending on the level. However, there is a tendency for the final topics to have a much smaller number of answers than the first ones. At the same time, (Topic -1 outlier) is present in all variants of calculations for 24 or 25 Topics with or without outlier.

Themes/Topics	1_city_building_police_buildings)	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Total	tions_vs_assigned_topic
		-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23		
1, 416 Applying for sponsorship	65	137	4	1	0	0	1	1	13	2	0	185	0	2	0	0	0	3	1	0	1	0	0	0	0	416	out16%-33%-44%
2, 511 Attending a seminar on stress r	0	285	225	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	511	56%-44%
3,619 Choosing a renewable energy so	0	0	0	254	0	1	362	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	619	41%-58%
4, 1216 Comparing two demographic g	0	5	1	631	0	579	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1216	52%-48%
5,296 Comparing two online retailers	2	1	0	188	0	85	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	296	64%-29%
6,449 Covering a news story	143	4	15	7	0	2	0	5	3	0	9	61	1	8	1	0	172	1	0	1	2	0	0	0	14	449	out32%-14%-38%
7,118 Delivering a punch line	40	1	1	0	0	0	0	0	0	0	0	74	1	0	0	0	0	0	0	0	1	0	0	0	0	118	out34%-63%
8,1132 Giving advice about budgeting	1	2	572	555	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1132	49%-51%
9,392 Hosting a group of foreign buyer	1	1	0	0	0	0	0	0	0	390	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	392	99%
10,73 Interpreting a prophecy	1	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	68	0	0	0	0	73	93%
11,282 Making a movie	4	0	4	1	0	24	1	0	0	0	0	0	0	0	0	192	0	0	1	0	0	54	0	1	0	282	68%-19%
12,167 Making an educational produc	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	165	0	0	0	0	0	0	167	99%
13,256 Reaching your potential	1	0	254	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	256	99%
14,1668 Renting out a room	172	2	0	2	901	0	0	589	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1668	54%-35%
15,459 Talking a friend out of a risky a	0	4	146	1	0	1	0	0	307	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	459	32%-67%
16,751 Writing a brochure for a muse	71	1	74	1	0	0	0	0	1	0	0	327	0	0	223	0	0	0	0	0	0	53	0	0	0	751	44%-30%
17,2441 Writing a campaign speech	9	1381	1038	4	0	1	2	0	0	0	0	1	1	0	2	0	0	0	0	1	0	0	0	0	1	2441	57%-43%
18,554 Writing a rejection letter	5	171	71	306	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	554	31%-13%-55%
19,128 Writing a wedding toast	0	0	0	0	0	1	0	0	127	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	128	99%
20,943 Writing about a disaster relief	63	442	15	8	3	3	1	1	0	2	343	1	0	0	0	61	0	0	0	0	0	0	0	0	0	943	47%-36%
21,81 Writing about future lifestyles	1	0	1	78	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	81	96%
22,706 Writing advertising copy	108	2	0	3	0	1	292	0	0	0	0	2	0	296	2	0	0	0	0	0	0	0	0	0	0	706	out15%-41%-42%
23,316 Writing an article about NLP te	22	2	2	1	0	182	0	0	0	1	0	1	0	1	0	0	0	0	0	76	0	0	0	0	28	0	316	58%-24%-9%
24,274 Writing an article about a sup	56	9	6	27	0	1	0	4	3	1	10	15	2	6	86	0	4	0	0	2	0	0	30	0	12	274	out20%-31%-11%
Total	765	2451	2429	2073	904	881	659	601	455	397	363	340	334	314	304	285	176	171	78	71	58	53	31	28	27	14248	
% Outliers	5,4																										

Figure 10: Correlation matrix for level C1: 24 Themes EFCAMDAT vs 24 Topics BERT.

Regarding the comparison of the 24 and 25 Topics, despite the fact that the 24 Topics do not correspond in number to the EFCAMDAT questions, as Topic -1 includes 'outliers' that are not included in the themes, the 24 Topics have a clearer correlation with 4 Topics keywords and the wording of the Themes.

For example, at the C1 level, in question *17.2441, Writing a campaign speech*, the keywords include the word vote 0 (0_student_council_vote_student_president), which corresponds to the idea of a campaign speech in the question, but in the variant of the distribution on 25 Topics, this word is no longer in the Top 4 words. It appears at the 5th place in the Top 10 keywords [‘*school’, ‘students’, ‘student’, ‘president’, ‘vote’, ‘council’, ‘thank’, ‘best’, ‘better’, ‘university’*].

The trend of a clearer correlation between Themes (questions) and Topics, when divided into 24, is also observed at all levels.

6. Discussions

The unequal distribution of topics can be explained by the fact that students with a lower level of language learning use fewer words when expressing their ideas. With fewer words, we obtain a smaller volume of lexical and semantic fields and a smaller number of Topics.

We assume that this also corresponds to the very idea of BERT: the less diverse the vectors, the more homogeneous the clusters.

7. Conclusions

This research explored the potential of BERTopic to assess semantic similarity between questions and answers in the EFCAMDAT corpus. By treating questions as topics and student answers as corresponding texts, we aimed to determine whether answers maintain a semantic correlation with their questions and how this correlation varies across language proficiency levels.

Our analysis revealed that BERTopic can effectively identify semantic links between questions and answers, with keyword analysis confirming a meaningful correlation. However, our results also indicate that topic modelling methods require further refinement to improve precision in determining semantic similarity. The comparison with the benchmark 20NewsGroups dataset demonstrated that topic distribution plays a crucial role in ensuring meaningful clustering. Specifically, our findings suggest that the optimal topic distribution for EFCAMDAT is 25 topics for A1-C1 levels and 9 topics for C2 levels, allowing a clearer mapping between student answers and the corresponding questions.

Beyond its immediate findings, this study provides a foundation for further research into text meaning. The ability to extract core semantic units from texts of varying length has broader implications for transfer learning in low-resource languages, such as Ukrainian, as well as in domain-specific applications. Future work should focus on refining topic modelling approaches to enhance their ability to capture nuanced semantic relationships and to reduce classification errors.

By advancing methods for semantic similarity detection, this research contributes to the broader field of computational linguistics and NLP, offering new insights into how topic modelling can be leveraged for analysing texts.

Acknowledgements

The research study depicted in this paper is funded by the program PAUSE ANR (the French National Research Agency) associated with the project ANR-17-CE19-0016 CLEAR (Communication, Literacy, Education, Accessibility, Readability).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. Bordes, X. Glorot, J. Weston, Joint learning of words and meaning representations for open-text semantic parsing, in: International Conference on Artificial Intelligence and Statistics, 2012.

[2] W. Gomaa, A. Fahmy, A survey of text similarity approaches, International Journal of Computer Applications 68 (2013) 13–18.
[3] H. T. Nguyen, P. H. Duong, E. Cambria, Learning short-text semantic similarity with word embeddings and external knowledge sources, Knowledge-Based Systems 182 (2019) 104842.
[4] M. H. Nguyen, D. Q. Tran, Estimation in semantic similarity of texts, Journal of Information Science and Engineering 37 (2021) 617–633. doi:10.6688/JISE.202105_37(3).0008.
[5] N. Khairova, Y. Kupriianov, A. Vorzhevitina, O. Shanidze, Models for effective categorization and classification of texts into specific thematic groups, in: CLW-2024: Computational Linguistics Workshop at 8th Int. Conf. on Computational Linguistics and Intelligent Systems (CoLInS-2024), Vol. 4, CEUR-WS, Lviv, Ukraine, 2024, pp. 37–49.
[6] Y.-S. Lin, Y. Jiang, S.-J. Lee, A similarity measure for text classification and clustering, IEEE Trans. on Knowledge and Data Engineering 26 (2014) 1575–1590. doi:10.1109/TKDE.2013.19.
[7] A. Kutuzov, M. Kopotev, T. Sviridenko, L. Ivanova, Clustering comparable corpora of Russian and Ukrainian academic texts: Word embeddings and semantic fingerprints, arXiv preprint arXiv:1604.05372 (2016).
[8] S. Ruder, A. Søgaard, I. Vulic, Unsupervised cross-lingual representation learning, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2019.
[9] F. Remy, P. Delobelle, H. Avetisyan, A. Khabibullina, M. de Lhoneux, T. Demeester, Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of LLMs for low-resource NLP, in: Proceedings of the Conference On Language Modeling, 2024.
[10] S. Harispe, S. Ranwez, J. Montmain, Semantic similarity from natural language and ontology analysis, Springer Nature, 2022.
[11] Y. Feng, E. Bagheri, F. Ensan, J. Jovanovic, The state of the art in semantic relatedness: a framework for comparison, The Knowledge Engineering Review 32 (2017) 1–30. doi:10.1017/S0269888917000029.
[12] P. B. Andersen, A theory of computer semiotics: semiotic approaches to construction and assessment of computer systems, Vol. 3, Cambridge University Press, 1990.
[13] H. Jackson, E. Zé Amvela, Words, Meaning, and Vocabulary, Continuum, 2000.
[14] M. Vakulenko, Semantic comparison of texts by the metric approach, Digital Scholarship in the Humanities 38 (2) (2022) 766–771. doi:10.1093/llc/fqac059.
[15] A. Akmajian, R. Demers, A. Farmer, R. Harnish, Linguistics: An Introduction to Language and Communication, MIT Press, 2001. doi:10.7551/mitpress/4252.001.0001.
[16] I. Mel'čuk, J. Milićević, An Advanced Introduction to Semantics: A Meaning-Text Approach, Cambridge University Press, 2020.
[17] A. Conneau, K. Khandelwal, et al., Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116v2 (2020).
[18] M. Pikuliak, M. Šimko, M. Bieliková, Cross-lingual learning for text processing: A survey, Expert Systems with Applications 165 (2021). doi:10.1016/j.eswa.2020.113765.
[19] M. Vakulenko, Deep contextual disambiguation of homonyms and polysemants, Digital Scholarship in the Humanities (2022). doi:10.1093/llc/fqac081.
[20] M. R. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, arXiv preprint arXiv:2203.05794 (2022).
[21] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, ... I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, arXiv preprint arXiv:2412.13663 (2024).
[22] G. J. Eertzen, A. T. Lexopoulou, A. Korhonen, Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCAMDAT), in: 31st Second Language Research Forum (SLRF), 2013.
[23] Council of Europe, A Common European Framework of Reference for Languages: Learning, Teaching, Assessment, Strasbourg, 2001.

[24] Y. Huang, J. Geertzen, R. Baker, A. Korhonen, T. Alexopoulou, The EF Cambridge Open Language Database (EFCAMDAT): Information for Users, University of Cambridge and EF Education First, 2017.
[25] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993–1022. doi:10.1162/jmlr.2003.3.4-5.993.
[26] D. M. Blei, J. D. McAuliffe, Supervised Topic Models, in: Advances in Neural Information Processing Systems, 2010.
[27] R. Egger, J. Yu, A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts, Frontiers in Sociology 7 (2022).
[28] O. Babalola, B. Ojokoh, O. Boyinbode, Comprehensive Evaluation of LDA, NMF, and BERTopic's Performance on News Headline Topic Modeling, Journal of Computing Theories and Applications 2 (2024) 268–289. doi:10.62411/jcta.11635.
[29] X. Wu, T. Nguyen, D. Zhang, W. Y. Wang, A. T. Luu, FASTopic: Pretrained Transformer is a Fast, Adaptive, Stable, and Transferable Topic Model, Advances in Neural Information Processing Systems 37 (2025) 84447–84481.

A. Online Resources

[30] A. Michalak, NLP research on the EFCAMDAT dataset, GitHub, URL: https://github.com/amichw/EFCAMDAT.
[31] EF Education First, Learn English online, EF English Live, URL: https://englishlive.ef.com/en-us/learn-english-online.
[32] M. Grootendorst, BERTopic documentation, URL: https://maartengr.github.io/BERTopic/index.html.
[33] Hugging Face, 20 newsgroups fixed dataset, URL: https://huggingface.co/datasets/rungalileo/20_Newsgroups_Fixed.
[34] Galileo, Improving your ML datasets with Galileo (Part 1), URL: https://www.galileo.ai/blog/improving-your-ml-datasets-with-galileo-part-1.
[35] R. Řehůřek, P. Sojka, Gensim: Topic modelling for humans, PyPI, URL: https://pypi.org/project/gensim/.
[36] AnswerDotAI, ModernBERT-base, Hugging Face, URL: https://huggingface.co/answerdotai/ModernBERT-base.