information

MDPI

Article

Topic Classification of Interviews on Emergency Remote Teaching

Spyridon Tzimiris *, Stefanos Nikiforos, Maria Nefeli Nikiforos, Despoina Mouratidis and Katia Lida Kermanidis

Humanistic and Social Informatics Laboratory, Department of Informatics, Ionian University, 49132 Corfu, Greece; nikiforos@ionio.gr (S.N.); nefeli.nikiforos@ionio.gr (M.N.N.); dmouratidi@ionio.gr (D.M.); kerman@ionio.gr (K.L.K.)

* Correspondence: c20tzim@ionio.gr

Abstract: This study explores the application of transformer-based language models for automated Topic Classification in qualitative datasets from interviews conducted in Modern Greek. The interviews captured the views of parents, teachers, and school directors regarding Emergency Remote Teaching. Identifying key themes in this kind of interview is crucial for informed decision-making in educational policies. Each dataset was segmented into sentences and labeled with one out of four topics. The dataset was imbalanced, presenting additional complexity for the classification task. The GreekBERT model was fine-tuned for Topic Classification, with preprocessing including accent stripping, lowercasing, and tokenization. The findings revealed GreekBERT's effectiveness in achieving balanced performance across all themes, outperforming conventional machine learning models. The highest evaluation metric achieved was a macro-F1-score of 0.76, averaged across all classes, highlighting the effectiveness of the proposed approach. This study contributes the following: (i) datasets capturing diverse educational community perspectives in Modern Greek, (ii) a comparative evaluation of conventional ML models versus transformer-based models, (iii) an investigation of how domain-specific language enhances the performance and accuracy of Topic Classification models, showcasing their effectiveness in specialized datasets and the benefits of fine-tuned GreekBERT for such tasks, and (iv) capturing the complexities of ERT through an empirical investigation of the relationships between extracted topics and relevant variables. These contributions offer reliable, scalable solutions for policymakers, enabling data-driven educational policies to address challenges in remote learning and enhance decision-making based on comprehensive qualitative evidence.

Keywords: topic classification; GreekBERT; Greek; interviews

Academic Editors: Barbara Pes and Andrea Loddo

Received: 18 February 2025
Revised: 17 March 2025
Accepted: 19 March 2025
Published: 21 March 2025

Citation: Tzimiris, S.; Nikiforos, S.; Nikiforos, M.N.; Mouratidis, D.; Kermanidis, K.L. Topic Classification of Interviews on Emergency Remote Teaching. *Information* **2025**, *16*, 253. https://doi.org/10.3390/info16040253

1. Introduction

Natural Language Processing (NLP) is a research field dedicated to creating techniques that allow computers to interpret and process human language [1]. One of the significant challenges in NLP is handling the enormous amount of text generated daily. Text classification, topic identification, topic modeling, and theme categorization refer to the process of categorizing unstructured text data into predefined themes or topics [2]. These terms are often used interchangeably in the literature to describe similar methodologies aimed at organizing and analyzing textual data for thematic content. Topic Classification (TC) is a core task within NLP, which categorizes text into specific themes based on content, enabling automated systems to retrieve relevant information efficiently from text corpora and sup-

port more informed decision-making for organizations, such as companies, governments, and research institutions [3–5].

Automated TC techniques are crucial because manually tagging large text collections is time- and effort-consuming. TC involves using a labeled dataset, in which every text sample is given a thematic-topic category, to train a machine learning (ML) model. TC includes identifying topics in news articles and social media posts, assessing sentiment in product reviews, and detecting spam in emails [3,6,7].

Recent studies have distinguished two main approaches to TC: conventional ML models and deep learning (DL) techniques. Conventional ML models perform effectively; however, DL approaches, particularly transformer-based models, have achieved superior results in capturing topic-thematic elements with higher accuracy [4,8,9]. Transformer models, with their encoder–decoder structure and attention mechanisms, can process entire texts in parallel, as opposed to earlier DL models that processed texts sequentially. This parallel processing enables them to handle large amounts of data efficiently, and their performance often improves with larger datasets, allowing them to capture more nuanced patterns and context in the text [10,11].

However, while TC research has largely focused on high-resource languages like English, studies on low-resource languages such as Greek are limited [12]. Papantoniou and Tzitzikas [13] review NLP progress for the Greek language, focusing on key tasks like text classification and sentiment classification (SC). They highlight the challenges posed by limited Greek-specific datasets and models, stressing the need for more tailored resources to advance Greek NLP applications across sectors. Transformer-based models have shown promising results in Greek NLP tasks, though Greek-specific datasets with DL models are still limited, and modern applications of DL in Greek are underexplored.

Our aim was neither to introduce a new model nor to present a novel architecture or technical advancements but to compare and evaluate the performance of conventional ML models (e.g., XGBoost) and DL models (e.g., mBERT, XLM-R Greek, and GreekBERT) in the context of Greek NLP applications, focusing on their evaluation using Greek datasets to provide insights into their relative effectiveness and limitations in low-resource language scenarios. This study aims to bridge these research gaps by introducing a new Modern Greek dataset and assessing the benefits of transformer-based models for Greek TC. Three original datasets collected from school directors, parents, and teachers, focusing on their perspectives regarding Emergency Remote Teaching (ERT) in Greece, are used in this study.

ERT is an unplanned and not well-structured approach to instruction, implemented as a temporary response in emergent situations, such as the COVID-19 pandemic. The term, as defined by Hodges et al. [14], distinguishes this form of teaching from intentionally designed and well-structured online learning.

Key contributions of this research include the following: (i) creating datasets capturing diverse educational community perspectives in Modern Greek, (ii) demonstrating how transformer-based models like GreekBERT effectively classified thematic content from interview datasets, and providing a comparative analysis with conventional ML models highlighting their strengths and limitations in handling non-English corpora, (iii) investigating how domain-specific language enhances the performance and accuracy of TC models, showcasing their utility in specialized datasets, and (iv) capturing the complexities of ERT through an empirical investigation into the relationships between extracted topics and relevant variables. TC can support decision-making processes by helping policymakers and educators gain deeper insights into stakeholder perspectives, ultimately enabling more informed and effective responses to educational challenges.

Based on the relevant literature review, the following research questions came up:

1. How effectively can transformer-based models classify thematic content from Modern Greek interview datasets?
2. How do ML and DL models compare in their effectiveness for TC in multi-class and domain-specific datasets?
3. How do models pre-trained in Greek perform in TC tasks?

2. Related Work

2.1. NLP in Greek

Research on low-resource languages like Greek has made significant progress in the field of NLP. Vagelatos et al. [15] developed an NLP environment tailored for supporting educational games in Greek, emphasizing the creation of linguistic resources to enhance interactive learning experiences. Krasadakis et al. [16] conducted a survey examining challenges and progress in NLP, particularly in low-resource languages like Greek, with a focus on legal informatics. They highlighted issues such as limited annotated corpora, a lack of domain-specific tools, and the complexity of legal texts. By the examination of rule-based systems, machine learning models, and neural architectures, the authors underscored the potential of transfer learning and pre-trained language models, and proposed strategies to address these gaps and advance NLP in this domain. Finally, Papantoniou and Tzitzikas [17] provided a comprehensive review of NLP tools and applications for the Greek language, identified current gaps, and proposed future research directions. In their research, the authors examined classical, statistical, and neural network approaches to Greek NLP; they focused on tools like tokenizers and lemmatizers, along with annotated corpora for tasks such as sentiment analysis and machine translation. As the authors mentioned, despite Greek lagging behind in terms of the available methods, tools, and resources, there is a growing interest in advancing NLP for the language. The aforementioned studies highlighted the importance of developing NLP solutions to address the unique linguistic and resource constraints of the Greek language and provided a strong foundation for applications like TC.

2.2. Topic Classification Techniques in Modern Greek Datasets

TC plays a crucial role in understanding qualitative data across various fields. Fang et al. [18] investigated the use of NLP techniques to classify qualitative data from cancer patient interviews, particularly in identifying symptoms and quality-of-life (QoL) impacts. Applying models such as TF-IDF, GloVe, Recurrent Neural Networks (RNNs), and BERT, they found BERT to be especially effective for categorizing patient feedback. BERT achieved the highest performance with a mean ROC AUC of 0.94. Cheng et al. [19] developed a method to improve topic extraction from interview data by combining multiple topic modeling approaches. Their best result was achieved with the Multi-Scale Hybridized Topic Modeling (MSHTM) method. It effectively combined the strengths of Non-negative Matrix Factorization (NMF) for capturing broad topics and BERTopic for identifying fine-grained subtopics, resulting in a clear hierarchical structure. Their study showed that this method could identify both broad and specific themes effectively in large-scale interview transcripts, outperforming standard topic modeling techniques. Liu and Sun [20] applied Large Language Models (LLMs) such as GPT-4 to classify themes from K-12 education policy stakeholder interviews. Their approach demonstrated a high level of accuracy, achieving 0.9 and alignment with human-coded thematic analysis, indicating that LLMs can effectively capture complex themes in interview data.

The application of TC in Modern Greek has grown significantly in recent years, driven by advancements in NLP. In the legal domain, Papaloukas et al. [21] introduced a multi-

granular Topic Classification system tailored for Greek legislation. Their work demonstrated the potential of NLP to enhance legal informatics and streamline the analysis of legislative texts, offering valuable tools for legal professionals. Beyond the legal sector, significant efforts have been made to explore TC and sentiment analysis social media content in Greek. Their Greek legislation Legal TC system achieved an F1-score of 0.89 using GREEK-LEGAL-BERT. Mastrokostas et al. [22] focused on classifying topics within Greek Reddit discussions, showcasing that TC techniques can uncover emerging themes and trends in user-generated content. The TC system for Greek Reddit posts achieved an F1-score of 0.79 with the use of GreekBERT. Similarly, Alexandridis et al. [23] conducted a comprehensive survey on sentiment analysis and opinion mining in Greek social media, revealing the challenges and progress in analyzing public opinions. The authors compared different approaches, and found that ML and DL methods consistently performed better than lexicon-based techniques in analyzing sentiment related to political, social, and economic topics. They also discussed the limited availability of annotated datasets, noting that most existing resources originated from platforms like Twitter and Facebook and were often tailored to specific domains such as politics and economics. Experimental results from their dataset showed that classifiers, such as Naive Bayes, Random Forests, SVMs, Logistic Regression, and Deep Feed-Forward Neural Networks, outperformed word or sentence embedding methods like Word2Vec and GloVe, achieving over 0.8 accuracy. Pitenis et al. [24] addressed the issue of offensive language identification in Greek, providing critical insights into detecting harmful content through text classification methods. Using a dataset of Greek social media posts annotated for offensive content, the LSTM and GRU with Attention model achieved the best performance with a macro-F1-score of 0.89. The study underscored the need for language-specific resources to effectively tackle Offensive Language Identification (OLI) in underrepresented languages like Greek. Michailidis [25] conducted a comparative analysis of sentiment classification models applied to Greek reviews, evaluating traditional ML methods alongside advanced approaches such as artificial neural networks (ANNs), transfer learning (TL), and LLMs. The findings revealed that transformer-based models, particularly GreekBERT (0.96 accuracy) and GPT-4 (0.95 accuracy), significantly outperformed other conventional methods. LLMs have demonstrated remarkable performance across various NLP tasks, including text classification in specialized domains such as banking [26]. Recent advancements have also led to the development of the first open Greek LLM, Meltemi, which aims to enhance NLP applications for the Greek language [27]. Furthermore, multi-task learning approaches incorporating natural language explanations have been explored for sentiment analysis and offensive language identification in Greek social media, providing a more user-centric evaluation framework [28]. These contributions highlight the growing emphasis on domain-specific and language-specific adaptations of LLMs, reinforcing their applicability in real-world scenarios.

The impact of TC and sentiment analysis has also been evident during major societal events. For example, Kydros et al. [29] conducted sentiment and content analysis on Greek tweets during the COVID-19 pandemic, providing insights into shifts in public sentiment across different stages of the crisis, focusing on topics like health measures, public restrictions, and social behaviors. They collected thousands of tweets in Greek during the peak of the pandemic and used NLP techniques, including lexicon-based and ML approaches, to analyze the data. Additionally, in multilingual and cross-lingual contexts, Chalkidis et al. [30] introduced MultiEURLEX, a dataset designed for legal document classification, enabling TC applications across multiple languages, including Greek. Lastly, Nikiforos et al. [31] provided an extensive survey of Greek-language datasets and text mining applications on social web platforms. Their study emphasized the need for robust

and tailored NLP tools to address the unique linguistic challenges posed by the Modern Greek language.

Despite the existing literature on TC in Modern Greek datasets, no research has focused specifically on interview data related to ERT. Existing datasets related to ERT have primarily been explored using methods like descriptive, thematic, or content analysis [32–34]. Most studies that used TC tasks have explored applications in the legal domain or social media contexts. While those works highlighted the versatility of TC across various Greek datasets, they also revealed a significant gap in applying these techniques to interview-based data. This study aims to fill this gap by introducing novel, education-focused datasets and utilizing TC to analyze school stakeholder perspectives in ERT, revealing insights into this unexplored area.

2.3. Classification Models

Transformer-based models have made advancements in text classification tasks by outperforming conventional ML methods, particularly in understanding complex text, and have become the state-of-the-art models for numerous NLP tasks. BERT's bidirectional context understanding, achieved through pre-training on large-scale corpora, has made it a powerful tool for tasks like TC [35].

Several variations of BERT have been developed to enhance its performance or to address specific challenges [36]. RoBERTa (Robustly Optimized BERT Pre-training Approach) improves upon BERT by training on larger datasets with longer sequences, achieving better results in various tasks [37]. Similarly, DistilBERT offers a smaller and faster version of BERT, making it more practical for resource-constrained applications while maintaining competitive accuracy [38]. BERT has also achieved high performance in multi-class TC tasks, achieving high levels of accuracy. Its ability to fine-tune on specific tasks allows it to handle multi-class and multi-language datasets effectively [39,40]. As mentioned earlier, despite the proven effectiveness of these models, their application to Greek text bodies remains limited, indicating a clear research gap that this study aims to fill [41].

Classification models in Modern Greek have made progress, and different techniques have been used to address the linguistic challenges of the Greek language. Approaches that focus on conventional ML models, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have been widely used. They are effective for more simple classification tasks, but they often struggle with the rich morphology and syntax complexity of Modern Greek. Transformer-based architectures like BERT have set new benchmarks in TC for the Greek Language.

GreekBERT developed by Koutsikakis et al. [42] has enhanced classification performance by being pre-trained on Greek-specific corpora. The model was trained on a large Greek dataset compiled from three sources: (i) Wikipedia articles [43], (ii) European Parliament Proceedings documents [44], and (iii) the OSCAR dataset [45]. This pre-trained model can be fine-tuned by other researchers to perform better in tasks such as TC and Named Entity Recognition (NER) compared to earlier deep learning models. It was pre-trained using two main tasks: Masked Language Modeling and Next Sentence Prediction. This model leverages the bidirectional transformer architecture to capture the intricate syntactic and semantic relationships in Greek text, making it highly suitable for TC in datasets such as interview transcripts.

Athinaios et al. [46] developed Greek-Legal-BERT, a novel linguistic model based on BERT for NER in Greek legal corpora. This was pre-trained on a 5 GB dataset comprising the entirety of Greek legislation and evaluated against GREEK-BERT, achieved comparable results for NER. Additionally, recent works on Greek classification datasets, including comparisons with traditional ML models, DL models, and GREEK-BERT, demonstrated

that the two BERT-based models consistently delivered the best overall performance, underscoring their effectiveness in handling Greek legal texts [47,48].

Also, models like GreekBART, the first pre-trained sequence-to-sequence model, have been developed recently, specifically designed for the Greek language, making a significant contribution to Greek NLP. This model can be used for discriminative tasks like text classification and Topic Classification, thanks to its encoder. Additionally, unlike BERT-based models, its decoder enables fine-tuning for generative tasks, such as text summarization [49].

Recently, GreekT5, which is based on the T5 architecture, has shown great improvement in Greek NLP. Proposed by Giarelis et al. [50], GreekT5 outperforms GreekBART in text summarization by most evaluation metrics. It uses a multilingual Seq2Seq architecture that is pre-trained to support the Greek language and is further fine-tuned for news summarization on the GreekSUM dataset. This model is versatile too and can be applied for other tasks like TC.

Lastly, Papadopoulos et al. [51] introduced a semantic textual similarity model, developed using Sentence-Transformers, to generate Greek sentence embeddings for a variety of NLP tasks. Known as lighteternal/stsb-xlm-r-greek-transfer, this model has been integrated into FarFetched, an automated framework designed for validating claims on Greek online news platforms. XLM-R Greek, derived from XLM-RoBERTa, has been fine-tuned specifically for Greek Natural Language Inference (NLI) and zero-shot classification tasks. Developed by the Hellenic Army Academy and the Technical University of Crete, this model was trained on a combined Greek and English version of the AllNLI dataset. The model is available at Hugging Face Repository [52].

The ability of BERT models to process large volumes of Modern Greek-language text efficiently makes them a valuable tool for researchers seeking to analyze large volumes of qualitative data efficiently. The current study evaluates the performance of ML and DL models on interview datasets, regarding ERT, focusing on their effectiveness in automated TC for qualitative analysis in Modern Greek.

2.4. Topic Classification and Decision-Making Approaches

Recent research has shown the potential of TC in supporting decision-making processes across various domains. Cao et al. [53] proposed a large-group emergency decision-making method that leverages topic sentiment analysis to assess risks in group decisions, demonstrating how sentiment-informed Topic Classification can enhance emergency responses. Similarly, Ahne et al. [54] developed an interactive classification and topic discovery approach for diabetes-related biomedical literature, aimed at aiding clinical decision-making by improving literature exploration. Their findings highlight the value of topic-based classification systems in the healthcare sector, particularly in synthesizing vast textual resources for clinical insights. Additionally, Huang et al. [55] utilized a topic modeling approach based on Latent Dirichlet Allocation (LDA) for sentiment classification of crowdsourced participant reviews, illustrating the utility of topic models in gathering feedback for informed decision-making. Leveraging Large Language Models for TC has shown promising applications in the domain of public affairs, enabling more nuanced and accurate categorization of themes within complex social and policy-related datasets. Those studies highlight the role of TC in facilitating more accurate and timely decisions through the structured analysis of textual data.

3. Material and Methods

The shift towards ERT during the COVID-19 pandemic presented numerous challenges in the Greek educational system [56]. The purpose of the collection of interviews was

to gather qualitative insights into participants' experiences with ERT during the COVID-19 pandemic, aiming to explore key challenges, adaptations, and their overall impact on the educational process. The perspectives of stakeholders—parents, teachers, and school directors—are crucial for informing data-driven decisions in educational policies. Identifying key themes in such stakeholder interviews helps policymakers to address issues like lack of digital infrastructure, teaching quality, psychological impact on students, and difficulties that arose from diverse functional needs. However, manually analyzing qualitative datasets is time-consuming. To mitigate these challenges, automated TC techniques using NLP can provide valuable solutions.

Interviews are a common qualitative data collection method, particularly in social sciences and educational research, where open-ended responses provide insights into participants' experiences, perceptions, and attitudes [57]. Thematic analysis is one of the most commonly used methods for analyzing interview data, offering a flexible approach to identifying key themes or topics within the responses, a task traditionally conducted manually by the researchers [58]. However, as datasets grow in size, manual analysis and labeling becomes impractical, leading to the adoption of NLP techniques such as TC [59].

The present study employs qualitative datasets derived from semi-structured interviews to examine ERT experiences among three key groups: parents of students with functional diversity, school directors, and teachers. The collected data span various demographics and geographical regions, urban, suburban, and rural areas in both mainland and island regions, ensuring a comprehensive understanding. Interviews were conducted in compliance with ethical guidelines, and all participants in this research signed informed consent forms in adherence to GDPR standards. All datasets and protocols used are available under the Creative Commons Attribution Non-Commercial No Derivatives 4.0 International license (CC BY-NC-ND 4.0) [60].

3.1. Participants and Statistics

Data were collected through interviews lasting an average of 45 min each, conducted remotely or in person. The Parents of Students with Functional Diversity dataset (PSFD dataset) comprised views from 12 parents (9 males and 3 females) whose children, diagnosed with various functional diversities, attended ERT during the pandemic. The sampling was designed to represent diverse functional needs, geographical areas, and school grades. The interviews with parents revealed a range of functional diversities among their children, including speech disorders, physical disabilities, general learning difficulties (GLDs), dyslexia, developmental dyscalculia, attention deficit hyperactivity disorder (ADHD), vision disabilities, and aggressiveness. Inclusion criteria required that parents had been present alongside their children throughout the ERT experience and dealt with their children's challenges. This dataset contained 14,827 words of qualitative data, organized into 1019 segments for analysis. Each segment represented a period of dialogue categorized by topic-thematic relevance to the study's objectives.

The PSFD dataset (see Table 1) displayed the most considerable variability, with an average word count of 1236 and a range of 1173, supported by a higher standard deviation of 325.9. Figure 1 below presents a line chart visualizing the word count distribution for each parent, providing a graphical view of response lengths across this group.

The notable variation in word counts may reflect a wide range of engagement and perspectives, influenced by varying degrees of participation in school matters or diverse personal experiences.

The School Directors dataset (SCHD dataset) consisted of 15 interviews with school directors, who provided insight into the administration challenges faced during the ERT period. The sample of school directors included 10 males and 5 females. The sampling was

designed to represent diverse geographical areas (both mainland and islands), and years of service. The average age of school directors was 51.07 years, while their average years of service in this role was 5.87 years. Inclusion criteria required that school directors had been active throughout the ERT experience and dealt with the administration challenges, leadership decisions, and adaptations necessary to support ERT. This dataset contained 17,171 words of qualitative data, organized into 1107 segments for analysis.

Table 1. PSFD Dataset: interview statistics.

Interviewees	Word Count	Functional Diversity of Their Child
I-1	805	Speech Disorder (stuttering, dysarthria), Physical Disability
I-2	1348	General Learning Difficulties (GLDs)
I-3	1507	Attention Deficit Hyperactivity Disorder (ADHD)
I-4	1142	Dyslexia, Developmental Dyscalculia
I-5	1978	General Learning Difficulties (GLDs)
I-6	985	Dyslexia, Speech Disorder (stuttering)
I-7	1034	General Learning Difficulties (GLDs)
I-8	1481	General Learning Difficulties (GLDs)
I-9	918	Attention Deficit Hyperactivity Disorder (ADHD), Aggressiveness
I-10	1174	General Learning Difficulties (GLDs)
I-11	1412	Vision Disability
I-12	1043	Attention Deficit Hyperactivity Disorder (ADHD)
Total	14,827	
Mean	1236	
Median	1158	
Standard Deviation	325.9	
Range	1173	
Minimum	805	
Maximum	1978	

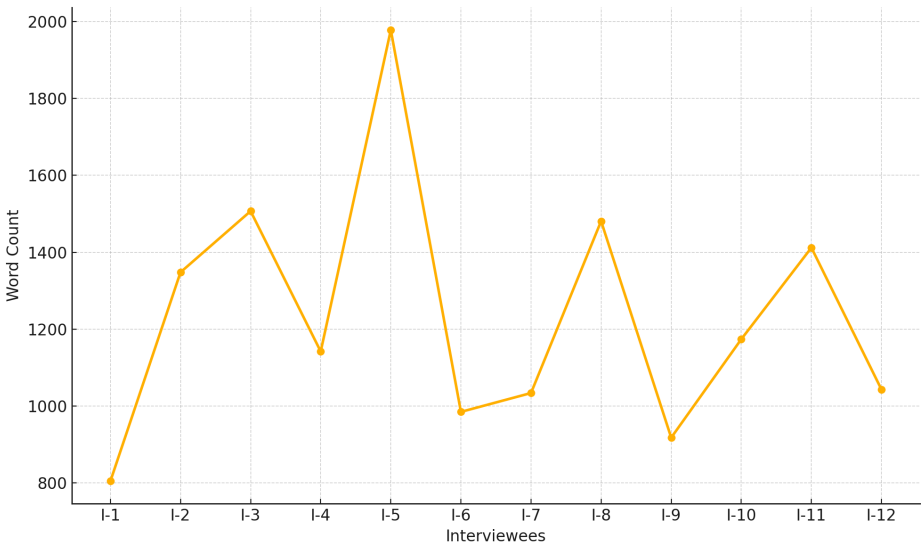

Figure 1. Line chart visualization of word count for each parent.

In the case of the SCHD dataset (see Table 2), the mean word count of 1145 and a more modest range of 569 suggests more consistency in response length. The lower standard deviation of 185.8 supports this pattern, pointing to greater uniformity in the school directors' responses. Figure 2 below shows the word count for each school director, highlighting the diversity in response lengths observed within this group.

Table 2. SCHD dataset: interview and demographic statistics.

Interviewees	Word Count	Age	Years of Service as School Directors
I-1	1338	50	5
I-2	1240	52	1
I-3	1410	62	12
I-4	899	45	7
I-5	1464	54	3
I-6	1080	40	4
I-7	1328	49	7
I-8	1098	57	5
I-9	1040	56	1
I-10	1153	55	8
I-11	1081	51	10
I-12	981	52	6
I-13	1243	50	2
I-14	921	45	15
I-15	895	48	2
Total	17,171		
Mean	1145		
Median	1098		
Standard Deviation	185.8		
Range	569		
Minimum	895		
Maximum	1464		

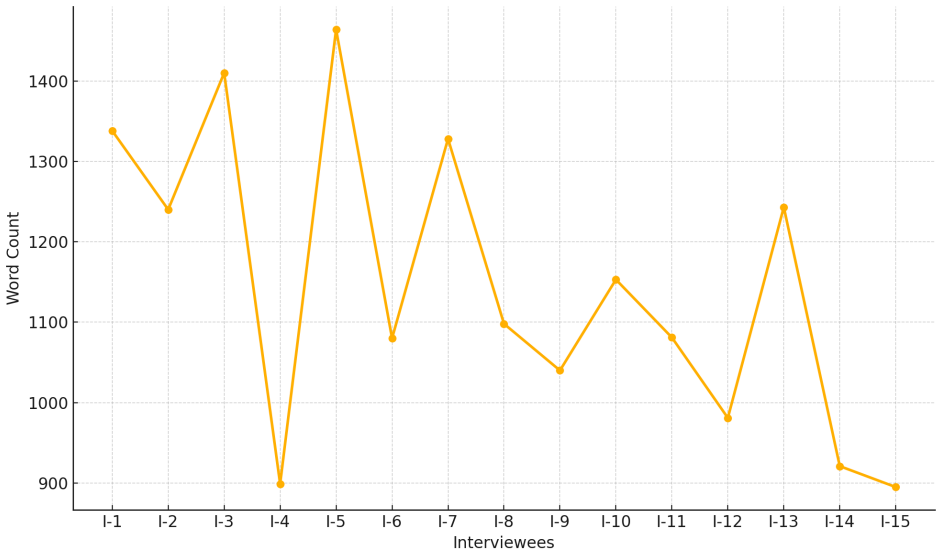

Figure 2. Line chart visualization of word count for each school director.

The uniformity in responses might suggest a shared set of perspectives or experiences, stemmed from similar administrative responsibilities or common challenges in their roles.

The Teachers dataset (TCH dataset) comprised 15 interviews with teachers (10 females and 5 males) involved in ERT process. The teacher interviews explored instructional challenges, resource allocation, and experiences with digital platforms during the shift to remote learning. Teachers were chosen from a cross-section of primary education schools, reflecting diverse instructional contexts and classroom settings. The sampling was designed to represent diverse geographical areas (both mainland and islands), and years of service. Inclusion criteria required that teachers had been teaching online lessons throughout

the ERT period. This dataset contained 28,014 words in Modern Greek, organized into 1224 segments for analysis.

For the TCH dataset (see Table 3), the average word count of 1868, with a range of 1229, suggested substantial variation in the depth of responses. This spread, from a minimum of 1239 to a maximum of 2468, indicates that while some teachers provided brief responses, others offered more detailed input. The following figure (Figure 3) presents the word count for each interviewee among teachers, illustrating the consistency and spread of responses within this cohort.

Table 3. TCH dataset: interview and demographic statistics.

Interviewees	Word Count	Age	Years of Service
I-1	1622	32	6
I-2	1239	45	20
I-3	1284	38	10
I-4	2236	50	25
I-5	2182	62	39
I-6	2420	55	30
I-7	1554	40	15
I-8	1904	48	22
I-9	2468	60	37
I-10	2000	52	28
I-11	1599	36	9
I-12	2114	46	21
I-13	1734	42	16
I-14	1939	49	23
I-15	1719	35	8
Total	28,014		
Mean	1868		
Median	1904		
Standard Deviation	377.1		
Range	1229		
Minimum	1239		
Maximum	2468		

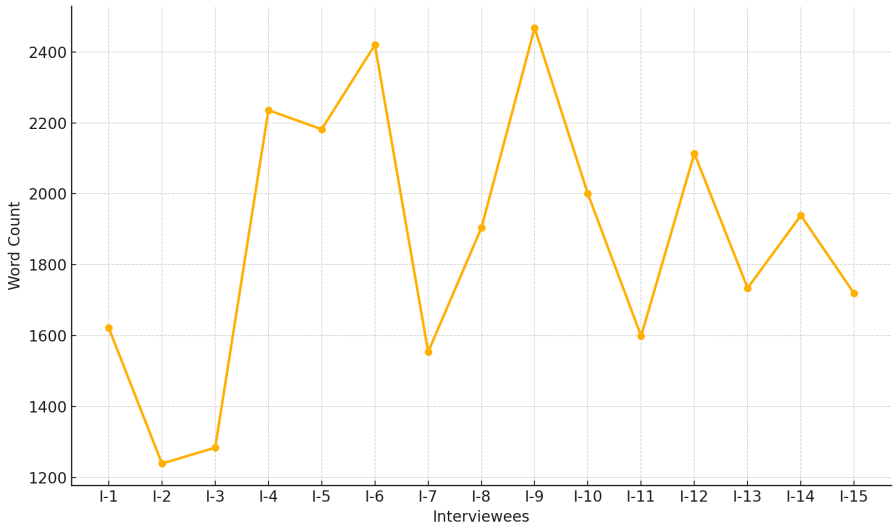

Figure 3. Visualization of word count for teachers.

The differences in responses could indicate varying levels of involvement or unique perspectives on the interview topics, emphasizing the diverse experiences among teachers.

These patterns illustrated unique characteristics for each group: teachers and parents exhibited more variability in responses, whereas school directors provided a more cohesive viewpoint. The PSFD dataset and SCHD dataset have also been used in previous studies [56,61]. However, these datasets were analyzed using different methodologies, such as thematic and linguistic analyses. The Teachers dataset is being presented in this study for the first time.

3.2. Data Preprocessing and Normalization

The qualitative data of interviews underwent extensive pre-processing to facilitate analysis:

1. Segmentation and Annotation: Interviews were segmented into four predefined topics. Class 1, regarding material and technical conditions, focused on factors related to infrastructure and resources. Class 2 addressed the educational dimension, emphasizing instructional practices and pedagogical strategies, while Class 3 captured the psychological–emotional dimension, reflecting the emotional and mental health aspects, and Class 4 dealt with learning difficulties and Emergency Remote Teaching, examining challenges faced by students with functional diversities during remote education settings. Based on the relevant literature in this field, these topics have been identified as crucial for understanding the key dimensions and the impact of the implementation of ERT [62,63].

The annotation process was conducted after segmentation. Datasets were manually annotated by two annotators at the sentence level, into the four predefined topics based on their semantic content. The first annotator (internal) was a member of the research team, closely involved with the project and possessing an in-depth understanding of the annotation guidelines. The second annotator (external) was a qualified researcher in the field, though not directly involved with the research project, who annotated the data based on provided guidelines through explicit examples. The inter-annotator agreement was 93%, indicating high consistency. In cases of annotator disagreement, a third annotator, who was also a qualified researcher in the field, was involved to resolve the ambiguity. Annotation ambiguity is a challenge that had to be resolved in order to ensure the execution of the process [64,65].

2. Data Preprocessing: No additional preprocessing steps were applied prior to utilizing the BERT models, as their architecture includes preprocessing capabilities. Experiments with additional preprocessing steps including stemming, stopwords removal, and lemmatization were conducted, but it was found that these processes were ineffective with the GreekBERT model, yielding worse results compared to when no preprocessing was applied. As mentioned by Haviana et al. [66], preprocessing techniques do not always improve the performance of the classification model. Their results also indicate that the use of stemming tends to decrease performance, a finding that aligns with the outcomes observed in our research.

3. Ethics and Informed Consent and Anonymization: Ethical considerations were prioritized throughout the research. Informed consent forms were collected from all the participants, outlining the purpose of data collection, secure storage of personal information, and exclusive research use of collected data [67]. In line with the GDPR requirements, all identifiable information was removed to ensure the privacy of the participants. Anonymity and confidentiality are essential to protect participants' privacy and maintain ethical integrity [68].

3.3. Licensing and Data Availability

The datasets generated and analyzed in this study are publicly available [69]. The repository provides open access to the data, including metadata, data files, and all associ-

ated protocols, supporting transparency, replicability, and the further exploration of ERT experiences during the COVID-19 pandemic.

4. Results

4.1. ML Models

Multiple ML and DL models were used in this study, including conventional models as well as latest transformer-based language models, in order to perform TC in the datasets. The selection of models aimed to achieve the highest performance. It was guided by the relevant literature regarding text classification tasks with Greek datasets [12]. These models have been widely used in similar studies and constitute established practices in the field. Our focus was on models that are either multilingual or pre-trained specifically in Greek, ensuring they were well suited to the scope and objectives of this study. This selection aligns with the requirements of the dataset and research goals. For each model, hyperparameter tuning was conducted to optimize performance. Models such as XGBoost, Generalized Linear Model (GLM), k-Nearest Neighbors (k-NN), Naive Bayes, and deep learning (H2O) were applied. For these models, textual data were transformed into numerical form using TF-IDF vectorization, which is a well-established baseline technique for text representation. This approach captures the statistical properties of the text, such as term frequency and document importance, but it does not retain semantic or contextual relationships. This limitation in representation likely contributed to the inferior performance of conventional machine learning methods compared to transformer-based models, which leverage the contextual embeddings. Preprocessing steps included converting text to lowercase, tokenization, and removing punctuation, as well as all Greek vowels (ά, έ, ί, ύ, ή, ό, ώ, ΐ, ΰ). Additionally, stopwords were filtered out, and tokens with fewer than four characters were removed to ensure a more informative and relevant feature set for text classification tasks. However, the results for GLM (F1-score: 0.33), k-NN (F1-score: 0.36), and Naive Bayes (F1-score: 0.38) were extremely low and close to random guessing, offering no meaningful insights. Therefore, we did not include them in the final analysis. Among these models, XGBoost emerged as the most effective, presenting better performance across evaluation metrics in the three datasets. As a result, XGBoost was selected as one of the four models chosen for the evaluation of the datasets.

Transformer-based architectures, including mBERT, XLM-R Greek, GreekBERT, were also incorporated. These models leveraged pre-trained embeddings, specifically optimized for Greek corpora, to capture the linguistic features of the text body. GreekBERT was selected because it is widely used for Greek text classification tasks. Although many transformer-based models exist, they are primarily designed for purposes other than text classification. GreekBERT's architecture and training corpus make it well suited for our specific classification needs, ensuring robust performance on Greek text. XLM-R-Greek and mBERT were also used, as they are state-of-the-art multilingual models. These models were selected to enable the comparison and performance assessment of multilingual models in Greek TC tasks. The evaluation of these models included metrics such as precision, recall, and F1-score, which were extracted using the classification report function by the Scikit-learn library.

To ensure the robustness of our results, we performed a stability analysis following the methodologies of [70,71]. Using repeated cross-validation and resampling, we evaluated the stability of key performance metrics for the models. This analysis confirmed that our findings are consistent and not overly sensitive to specific data splits, enhancing the reliability and reproducibility of the results.

After extensive testing with various hyperparameter configurations, including learning rate, batch size, number of epochs, regularization parameters (L1/L2), number of layers

or neurons, and dropout rate, the optimal settings for each model were identified and retained. These parameters, detailed in Appendix Table A1, were used to run each model five times, ensuring consistency, with the averaged results presented in the results section to ensure validity and reliability. This approach allowed us to fine-tune the models, aiming for the best performance on the dataset.

Figure 4 below shows the process followed for the TC.

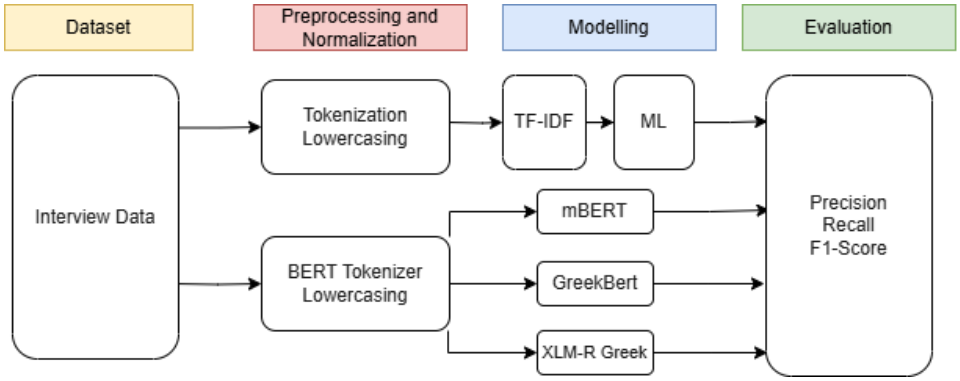

Figure 4. Workflow for Topic Classification.

4.2. Classification Performance

All three datasets (PSFD SCHD, and TCH datasets) using ML models for TC were evaluated in this study. The performance of these models were evaluated in classifying responses across distinct participant groups, and comparisons were conducted across the metrics of XGBoost, mBERT, XLM-R Greek, and GreekBERT. Table 4 presents the macro-averaged metrics per model for each dataset, highlighting the best performance. These comparisons offered valuable insights on model effectiveness and aided in selecting the most suitable approach for accurately classifying data.

Table 4. Performance metrics across datasets and models.

Dataset	Model	Precision	Recall	F1-Score
SCHD Dataset	XGBoost	0.52	0.43	0.44
	mBERT	0.67	0.71	0.68
	XLM-R Greek	0.71	**0.75**	0.72
	GreekBERT	**0.80**	0.73	**0.76**
PSFD Dataset	XGBoost	0.60	0.57	0.58
	mBERT	0.69	0.68	0.67
	XLM-R Greek	0.72	0.72	0.72
	GreekBERT	**0.78**	0.71	0.74
TCH Dataset	XGBoost	0.54	0.54	0.53
	mBERT	0.71	0.69	0.70
	XLM-R Greek	**0.80**	**0.79**	**0.79**
	GreekBERT	0.76	**0.78**	**0.76**

The results presented in this table represent the macro-averaged performance metrics for each dataset across different models. The best results per case are highlighted.

For the SCHD dataset, GreekBERT achieved the highest precision of 0.80 and the highest F1-score of 0.76. XLM-R Greek recorded the highest recall of 0.75, alongside a precision of 0.71 and an F1-score of 0.72. mBERT showed 0.68 F1-score, with its highest metric being 0.71 recall and 0.67 precision. XGBoost achieved its highest precision of 0.52, with 0.43 recall and 0.44 F1-score.

In the PSFD dataset, GreekBERT achieved the highest precision (0.78) and the highest F1-score (0.74), and 0.71 recall. XLM-R Greek recorded the highest recall of 0.72, along with 0.72 precision and 0.72 F1-score. mBERT showed 0.69 precision, which is its highest metric, along with a recall of 0.68 and 0.67 F1-score. XGBoost achieved its highest precision (0.60), with 0.57 recall and 0.58 F1-score.

In the TCH dataset, XLM-R Greek achieved the highest precision (0.80) and the highest F1-score (0.79), and 0.79 recall. GreekBERT recorded the highest recall (0.78), with 0.76 precision and 0.76 F1-Score. mBERT showed its highest precision of 0.71, 0.69 recall, and 0.70 F1-score. XGBoost achieved an F1-score of 0.53, its highest metrics being precision and recall, both at 0.54.

XGBoost demonstrated its strongest performance in the PSFD dataset, achieving its highest precision (0.60). In the SCHD dataset, XGBoost recorded a precision of 0.52, which was its highest metric in that dataset. In the TCH dataset, XGBoost reached its highest metric with an 0.53 F1-score, supported by precision and recall of 0.54.

mBERT performed best in the SCHD dataset, achieving its highest recall (0.71) and 0.68 F1-score. With the PSFD dataset, its strongest metric was precision of 0.69, with 0.67 F1-score. For the TCH dataset, mBERT's best result was precision of 0.71, 0.70 F1-score, and 0.69 recall.

XLM-R Greek achieved its strongest metrics in the TCH dataset, where it recorded the highest precision (0.80) and the highest F1-score (0.79). With the SCHD dataset, it attained the highest recall (0.75) and 0.72 F1-score. In the PSFD dataset, XLM-R Greek achieved balanced metrics with a precision, recall, and F1-score all at 0.72.

GreekBERT achieved its strongest 0.80 precision in the SCHD dataset, where it also recorded the highest F1-score (0.76). In the PSFD dataset, GreekBERT attained the highest precision (0.78) and an 0.74 F1-score. For the TCH dataset, GreekBERT achieved its highest recall (0.78), along with an 0.76 F1-score and precision.

Unlike Table 4, which emphasizes the comparison of models, Table 5 presents the macro-averaged performance metrics of various models across participant datasets, facilitating a comparative analysis of model performance for each dataset. This distinction was intentionally made to enhance clarity and support a more precise understanding of the comparative analyses.

Table 5. Performance metrics across models per participant dataset.

Model	Dataset	Precision	Recall	F1-Score
XGBoost	SCHD dataset	0.52	0.43	0.44
	PSFD dataset	0.60	0.57	0.58
	TCH dataset	0.54	0.54	0.53
mBERT	SCHD dataset	0.67	0.71	0.68
	PSFD dataset	0.69	0.68	0.67
	TCH dataset	0.71	0.69	0.70
XLM-R Greek	SCHD dataset	0.71	**0.75**	0.72
	PSFD dataset	0.72	0.72	0.72
	TCH dataset	**0.80**	**0.79**	**0.79**
GreekBERT	SCHD dataset	**0.80**	0.73	**0.76**
	PSFD dataset	**0.78**	0.71	0.74
	TCH dataset	**0.76**	**0.78**	**0.76**

The results presented in this table represent the macro-averaged performance metrics across all four classes within the datasets.

The performance metrics revealed differences among the models when applied to the datasets. XGBoost achieved its highest performance in the PSFD dataset, with 0.60 precision.

mBERT presented its strongest performance in the SCHD dataset, with 0.71 recall. XLM-R Greek demonstrated its best 0.80 precision in the TCH dataset. GreekBERT attained its highest precision of 0.80 in the SCHD dataset.

While the highest metric was achieved by GreekBERT and XLM-R Greek, both with a precision of 0.80 in the TCH dataset, GreekBERT showed consistently strong performance across the datasets, highlighting the benefits of language-specific pre-trained models for Greek text classification.

The highest precision was achieved by GreekBERT in the SCHD dataset, recording a value of 0.80. GreekBERT also performed well with 0.78 precision in the PSFD dataset, while XLM-R Greek matched the highest precision (0.80) in the TCH dataset. mBERT achieved its highest precision of 0.71 in the TCH dataset, and XGBoost demonstrated its best precision (0.60) in the PSFD dataset.

XLM-R Greek achieved the highest recall in the SCHD dataset at 0.75 and in the PSFD dataset at 0.72. In the TCH dataset, GreekBERT recorded the highest 0.78 recall. mBERT achieved its highest recall of 0.71 in the SCHD dataset, and XGBoost showed a 0.57 recall in the PSFD dataset.

In the SCHD dataset, GreekBERT achieved the highest F1-score of 0.76, followed by XLM-R Greek with 0.72. In the PSFD dataset, GreekBERT led with an 0.74 F1-score, while XLM-R Greek achieved a score of 0.72. In the TCH dataset, XLM-R Greek attained the highest F1-score (0.79), while GreekBERT recorded 0.76. mBERT's highest F1-score was 0.70 in the TCH dataset, and XGBoost reached its best F1-score (0.58) in the PSFD dataset.

Since GreekBERT was the best performing model, the classification results used GreekBERT for each dataset are presented in Table 6. This table provides metrics for each class across the three groups.

Table 6. GreekBERT results per dataset and per class.

SCHD Dataset	**Precision**	**Recall**	**F1-Score**
Class 1	0.80	**0.87**	0.84
Class 2	**0.54**	0.45	0.49
Class 3	**0.91**	0.77	0.83
Class 4	0.93	0.82	**0.88**
PSFD Dataset	**Precision**	**Recall**	**F1-Score**
Class 1	**0.88**	0.74	0.81
Class 2	0.68	**0.74**	0.71
Class 3	0.65	**0.77**	0.71
Class 4	**0.89**	0.59	0.71
TCH Dataset	**Precision**	**Recall**	**F1-Score**
Class 1	**0.87**	0.80	**0.83**
Class 2	0.69	0.72	0.71
Class 3	0.66	**0.84**	0.74
Class 4	**0.81**	0.74	0.77

This table presents precision, recall, and F1-score metrics for each class across directors, parents, and teachers, with the highest values in each metric per class highlighted. Class 1: Material and Technical Conditions, Class 2: Educational Dimension, Class 3: Psychological–Emotional Dimension, Class 4: Learning Difficulties and Emergency Remote Teaching.

In the SCHD dataset, Class 1 showed its strongest result in recall, achieving 0.87. Class 2 performed best in precision, reaching 0.54. Class 3 had its highest metric in precision at 0.91, while Class 4 achieved its best result in F1-score with 0.88.

For the PSFD dataset, Class 1 achieved its highest metric in precision with 0.88. Class 2 reached its best recall at 0.74. Class 3 also demonstrated its strongest result in recall, attaining 0.77. Class 4 exhibited its top precision with a value of 0.89.

In the TCH dataset, Class 1 achieved its strongest results in both precision and F1-score, with values of 0.87 and 0.83, respectively. Class 2 attained its highest metric in recall with 0.72. Class 3 recorded its best result in recall at 0.84, and Class 4 performed best in precision with a value of 0.81.

The confusion matrices in Figures 5–7 present GreekBERT's classification performance across the Directors, Parents, and Teachers datasets, respectively. These matrices highlight the model's ability to accurately classify each category within the data and provide insights into areas where misclassifications may occur, informing potential improvements for future applications.

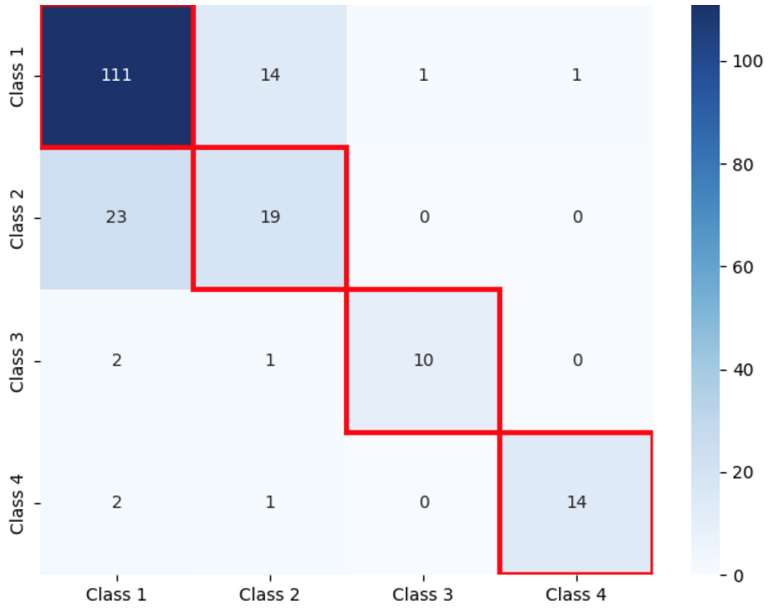

Figure 5. Confusion matrix: SCHD dataset. The diagonal red frames highlight the correctly classified instances.

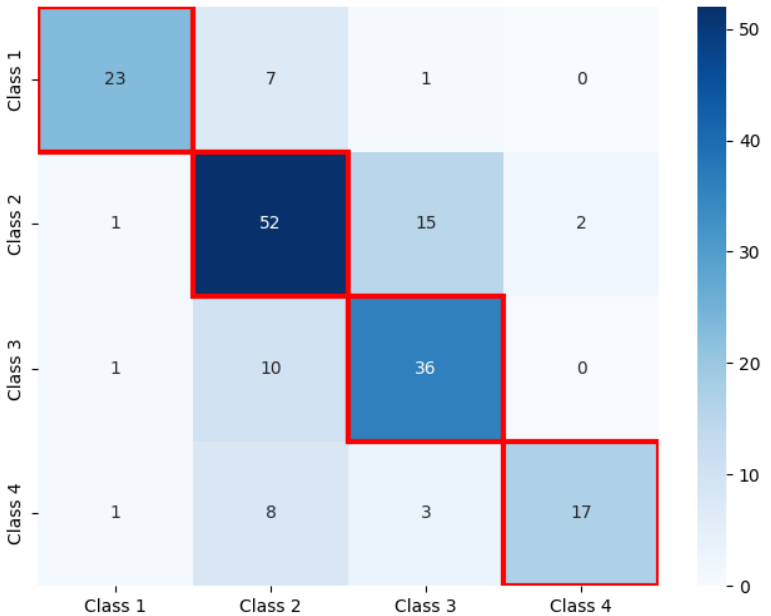

Figure 6. Confusion matrix: PSFD dataset. The diagonal red frames highlight the correctly classified instances.

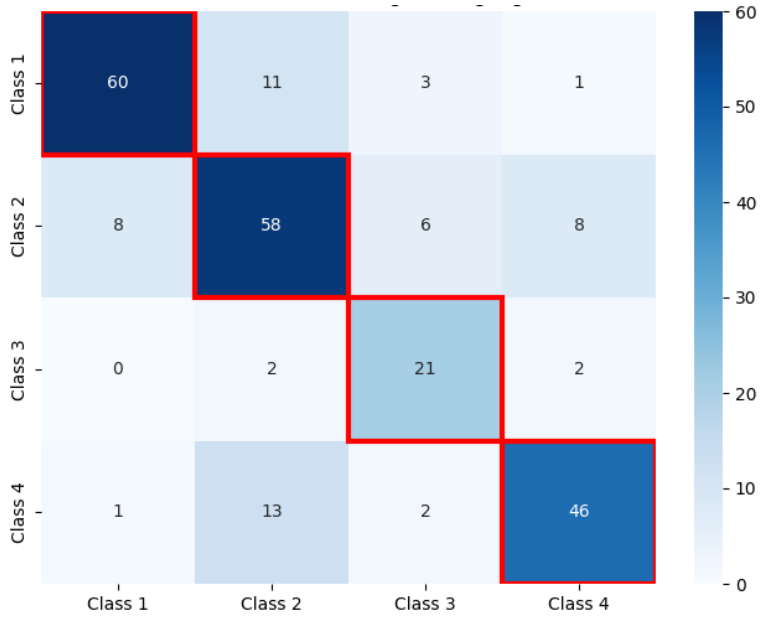

Figure 7. Confusion matrix: TCH dataset. The diagonal red frames highlight the correctly classified instances.

The computational experiments were performed on a system equipped with an 11th Gen Intel Core i7-11800H CPU with a base clock of 2.30 GHz, 16 GB of RAM, a 64-bit architecture, complemented by an NVIDIA GeForce GPU with 6 GB of VRAM.

4.3. Comparison

The comparison with relevant research focused on the general application of NLP for TC in Modern Greek data, the use of conventional machine learning methods and transformer-based models for text classification. It also addressed the implementation of TC approaches in interview data and the application of BERT models on Modern Greek corpora, aiming to provide a comprehensive and broad examination that aligned as closely as possible with the characteristics and context of the data utilized in the present study. This study achieved strong predictive metrics using GreekBERT, which closely matched the metrics reported in research employing similar methodologies. Comparing with relevant research helped to understand how the current work aligns with or differs from existing research, highlighting its unique contributions and validating its findings within a broader context (see Table 7).

Table 7. Comparison with relevant research.

Research	Data	Task	Classes	Model	Results
[22]	Greek Reddit posts	TC	10	GreekBERT	F1: 0.79
[21]	Greek legislation	Legal TC	50	GREEK-LEGAL-BERT	F1: 0.89
[25]	Greek reviews	SC	3	BERT, GPT-4	F1: 0.96
This paper	Greek Interviews	TC	4	GreekBERT	F1: 0.76

TC: Topic Classification, SC: Sentiment Classification, F1: F1-Score.

5. Discussion

The topics extracted in this study were directly connected to key aspects of ERT, including student engagement, technological challenges, teaching practices, and the psychological and emotional dimensions of the learning experience. This alignment emphasizes the relevance of our findings to the ERT context and reflects similar observations in previous

research [72,73]. By focusing on these specific dynamics, the study offers valuable insights into the challenges and patterns of remote teaching during emergency situations.The approach of this study has proven to be functional and effective in identifying meaningful patterns within the ERT context.

From the statistics of the interviews, as presented in Section 3.1, Participants and Statistics, the patterns observed illustrate the unique characteristics for each group. The word count analysis across the three groups reveals notable distinctions in response length and variability, suggesting different levels of engagement and perspectives. Teachers and parents exhibit more variability in responses, whereas school directors provide a more cohesive viewpoint. For the PSFD dataset, this significant spread in word counts may signify a broad spectrum of engagement and viewpoints, potentially shaped by differing levels of involvement in school affairs or varied personal experiences. This diversity in the parents' responses could indicate a range of perspectives on school-related topics, underscoring the different ways that parents perceive or interact with educational matters. For the SCHD dataset, this consistency could indicate a more aligned set of perspectives or experiences, potentially due to shared administrative roles or common challenges within their professional scope. For the TCH dataset, this variability may reflect diverse levels of engagement or distinct viewpoints on the interview topics, highlighting differences in experiences among the teachers.

This study introduced original interview datasets for TC. A range of models for multi-class TC were used, incorporating both ML classifiers and DL approaches. Experimental results revealed that ML models utilizing pre-trained transformer-based embeddings, such as XLM-R-Greek, achieved high performance. However, the fine-tuned GreekBERT model outperformed all other ML models, excelling in evaluation metrics, like precision, recall, and F1-score.

To evaluate the performance of the models for each dataset, confusion matrices were analyzed. The findings showed that the DL model had fewer false positives and false negatives, while the ML models had more false negatives.

GreekBERT demonstrated strong performance in TC across multi-class data, effectively predicting class labels and capturing themes related to ERT. The model outperformed conventional approaches, accurately differentiating topics and reflecting the unique perspectives of key educational stakeholders, including directors, parents, and teachers. Classification performance underscores the importance of using pre-trained models in the Greek language. These models effectively capture linguistic and contextual details, leading to the more accurate classification of domain-specific data as shown in other works [74,75].

The results showed that different models had varying levels of effectiveness across the datasets. GreekBERT consistently performed well, revealing its ability to handle language-specific tasks with a strong balance between precision and recall. XLM-R Greek also performed effectively, particularly in tasks requiring high recall, indicating its versatility and strength. mBERT presented moderate results across datasets, suggesting that while it was adaptable, it could benefit from further tuning for Greek-specific applications. XGBoost, although not as competitive overall, provided reasonable outcomes in certain datasets, showing that simpler models could still be useful in specific contexts.

While the highest metrics were achieved by the GreekBERT and the XLM-R Greek models with a precision of 0.80 in the Teachers dataset, GreekBERT showed better overall performance across datasets, highlighting the benefits of language-specific pre-trained models for Greek text classification, as they capture linguistic features more effectively, leading to higher performance across datasets.

The results reveal that GreekBERT performed strongly in identifying and classifying topics for Class 1 and Class 4, with Class 4 consistently showing the best overall outcomes

across the PSFD dataset, SCHD dataset, and TCH dataset. Class 1 topics were handled with high precision and consistency, particularly in the PSFD dataset and SCHD dataset. Class 2 showed lower overall performance, although the model managed to classify these topics effectively within the PSFD dataset and TCH dataset. This difference in performance occurred because the words used by the interviewees in Class 2 are commonly found in other classes as well. Frequently used words such as "child" ("παιδί"), "school" ("σχολείο"), "teacher" ("δάσκαλος"), and "lesson" ("μάθημα") likely caused confusion for the model, leading to lower performance in this class. For Class 3, the model excelled in classifying topics in the SCHD dataset while maintaining balanced performance in the TCH dataset. GreekBERT was especially accurate and consistent in classifying Class 4 topics, particularly in the SCHD dataset, while also showing strong sensitivity in detecting relevant responses in the TCH dataset.

The results revealed that as the experimentation incorporates more state-of-the-art models, particularly those pre-trained in Greek, such as GreekBERT and XLM-R Greek, the performance metrics, including precision, recall, and F1-score, presented an improvement.

GreekBERT revealed significant potential in aiding policymakers by transforming qualitative data into clear, data-driven insights. By categorizing interview statements from directors, parents, and teachers, the model identified key issues, challenges, areas for improvement, and stakeholder priorities, enabling rapid responses in critical situations such as ERT. For instance, it addressed specific challenges faced by the parents of students with functional diversities by achieving equitable access to resources and facilities for all students in inclusive education, while insights from teachers and directors informed targeted solutions for the learning process and the administration needs.

From the comparison with relevant research, the studies reviewed and highlighted different applications of NLP techniques for Modern Greek text classification, ranging from SC to Legal TC and social media TC. Compared to these works, the current study focused on classifying interview data and differentiated as the first to apply TC specifically to Greek interview data, addressing the complexities of this unique domain. GreekBERT, fine-tuned for this task, demonstrated its strength in handling context-specific content, achieving results that aligned with or exceeded those reported in other research. This work contributes significantly to the field by exploring an underrepresented area, providing a framework for future research on Greek interview corpora, and showcasing the adaptability of domain-specific transformers. Additionally, this study achieved comparable or superior macro-averaged performance metrics (macro-F1-score: 0.76) compared to relevant works. This distinction underscores the contribution of this work in addressing less explored data types while aligning closely with the methodologies and results of relevant research.

The model's high accuracy helps to highlight challenges and opportunities and guide the development of interventions that enhance ERT experiences.

Limitations

The Greek datasets used in this work reveal educational stakeholders' perspectives within the context of ERT. The dataset size may appear small for machine learning experiments. However, it consists of real humanitarian data collected under actual conditions, capturing essential linguistic and contextual details often missing from larger, more general corpora, providing valuable qualitative richness and domain-specific significance. Also, the limited availability of Greek NLP datasets with varied themes and writing styles restricts the development and evaluation of broader general-purpose TC models. One more notable limitation is the reliance on pre-trained models, which, while effective, could yield better results with further training on larger and more diverse datasets to capture linguistic expressions and domain-specific terminology in the Greek educational context.

Transformer-based models, like GreekBERT, trained on datasets like Wikipedia, European Parliament Proceedings, and OSCAR may struggle to fully capture the context-specific terminology when applied to domain-specific data. Additionally, the findings are shaped by the quality and representativeness of the datasets, which may limit the applicability of the results to broader populations or different contexts. Another limitation concerns the generalizability of the findings, with few pre-trained language models available for Modern Greek; further evaluations using future Greek-specific LMs and additional datasets are necessary to confirm the reliability of the results.

6. Conclusions and Future Work

Transformer-based architectures have significantly reshaped the field of TC, emphasizing the value of pre-trained embeddings. This study contributes to the field of automated analysis of educational data through the following: (i) the creation of datasets that capture the perspectives of school stakeholders in Modern Greek regarding ERT, (ii) a comparative analysis between conventional ML models and transformer-based approaches, (iii) empirical evidence demonstrating the benefits of fine-tuned GreekBERT for domain-specific applications, (iv) an empirical investigation into the relationships between extracted topics and relevant variables in the context of ERT, combined with the development of automated techniques for analyzing qualitative, education-related data, and (v) emphasizing how TC can support decision-making processes by helping policymakers and educators gain deeper insights into stakeholder perspectives, ultimately enabling more informed and effective responses to educational challenges. These contributions provide reliable tools for policymakers, enabling the formulation of data-driven educational policies that address the challenges of remote learning and enhance decision-making processes through qualitative data.

Future research could apply transformer-based models for topic-based sentiment analysis on the dataset. Understanding sentiment within specific topics would provide deeper insights into stakeholder experiences in ERT contexts. It could also explore additional participant groups and contexts to gain a broader understanding of educational challenges. Adding more datasets for fine-tuning the model could improve its adaptability and performance. GreekBERT's ability to generalize across participant groups suggests scalability to similar datasets, making it valuable for future studies in emergency education or other scenarios requiring swift policy responses. By leveraging GreekBERT's structured analysis capabilities, policymakers can develop tailored strategies, and create adaptive policies that effectively address stakeholder needs, ensuring a responsive approach to educational challenges.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision, and project administration were collaboratively performed by all authors. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This research required ethical approval and was conducted in accordance with institutional ethical guidelines and approvals. It was approved by the Ionian University Ethics Committee, under protocol code 869/07-03-2023. All ethical standards were strictly followed to ensure participant confidentiality and the responsible handling of data throughout the research process.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study. Written informed consent forms outlined the purpose of data collection, the secure storage of personal information, and exclusive research use, in compliance with ethical standards.

Data Availability Statement: The datasets generated and analyzed in this study are publicly available at https://hilab.di.ionio.gr/index.php/en/datasets/ (accessed on 10 February 2025). This repository ensures open access to the data, allowing researchers and practitioners to explore and build upon the findings presented in this study. The dataset repository includes metadata, data files, and all associated protocols, in compliance with the Creative Commons Attribution Non-Commercial No Derivatives 4.0 International License, accessible at https://creativecommons.org/licenses/by-nc-nd/4.0/ (accessed on 10 February 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Parameters by model.

Model	Parameters
XGBoost	Learning Rate: 0.4 Max Depth: 6 Number of Estimators: 50 Subsample Ratio: 1.0 Gamma: 0.0 Early Stopping: None Objective Function: Multi-Class Classification Test Size: 0.2 Feature Extraction:TF-IDF
mBERT	Learning Rate: 5×10^{-5} Batch Size: 8 Maximum Sequence Length: 128 Number of Epochs: 10 Dropout: 0.1 Test Size: 0.2
XLM-R Greek	Learning Rate: 5×10^{-5} Batch Size: 8 Maximum Sequence Length: 128 Number of Epochs: 10 Dropout: 0.1 Test Size: 0.2
GreekBERT	Learning Rate: 5×10^{-5} Batch Size: 8 Maximum Sequence Length: 128 Number of Epochs: 10 Dropout: 0.1 Test Size: 0.2

References

1. Khurana, D.; Koli, A.; Khatter, K.; Singh, S. Natural language processing: State of the art, current trends and challenges. *Multimed. Tools Appl.* **2023**, *82*, 3713–3744. [PubMed]
2. Abdelrazek, A.; Eid, Y.; Gawish, E.; Medhat, W.; Hassan, A. Topic modeling algorithms and applications: A survey. *Inf. Syst.* **2023**, *112*, 102131.
3. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–40. [CrossRef]
4. Li, Q.; Peng, H.; Li, J.; Xia, C.; Yang, R.; Sun, L.; Yu, P.S.; He, L. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol. (TIST)* **2022**, *13*, 1–41.
5. Gasparetto, A.; Marcuzzo, M.; Zangari, A.; Albarelli, A. A survey on text classification algorithms: From text to predictions. *Information* **2022**, *13*, 83. [CrossRef]

6. Satu, M.S.; Khan, M.I.; Mahmud, M.; Uddin, S.; Summers, M.A.; Quinn, J.M.; Moni, M.A. TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets. *Knowl.-Based Syst.* **2021**, *226*, 107126. [PubMed]
7. Nikiforos, M.N.; Deliveri, K.; Kermanidis, K.L.; Pateli, A. Vocational Domain Identification with Machine Learning and Natural Language Processing on Wikipedia Text: Error Analysis and Class Balancing. *Computers* **2023**, *12*, 111. [CrossRef]
8. Jelodar, H.; Wang, Y.; Orji, R.; Huang, S. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE J. Biomed. Health Informat.* **2020**, *24*, 2733–2742.
9. Lavanya, P.; Sasikala, E. Deep learning techniques on text classification using Natural language processing (NLP) in social healthcare network: A comprehensive survey. In Proceedings of the 2021 3rd International Conference on Signal Processing and Communication (ICPSC), Coimbatore, India, 13–14 May 2021; pp. 603–609.
10. Palani, S.; Rajagopal, P.; Pancholi, S. T-BERT–Model for Sentiment Analysis of Micro-blogs Integrating Topic Model and BERT. *arXiv* **2021**, arXiv:2106.01097.
11. Omar, A.; Mahmoud, T.M.; Abd-El-Hafeez, T.; Mahfouz, A. Multi-label arabic text classification in online social networks. *Inf. Syst.* **2021**, *100*, 101785. [CrossRef]
12. Liapis, C.M.; Kyritsis, K.; Perikos, I.; Spatiotis, N.; Paraskevas, M. A Hybrid Ensemble Approach for Greek Text Classification Based on Multilingual Models. *Big Data Cogn. Comput.* **2024**, *8*, 137. [CrossRef]
13. Papantoniou, K.; Tzitzikas, Y. NLP for the Greek language: A brief survey. In Proceedings of the 11th Hellenic Conference on Artificial Intelligence, New York, NY, USA, 2–4 September 2020; pp. 101–109.
14. Hodges, C.B.; Moore, S.; Lockee, B.B.; Trust, T.; Bond, M.A. The difference between emergency remote teaching and online learning. *Educ. Rev.* **2020**, *27*, 1–9.
15. Vagelatos, A.; Stamatopoulos, J.; Fountana, M.; Gavrielidou, M.; Tsalidis, C. Natural Language Processing Environment to Support Greek Language Educational Games. In *Interactive Mobile Communication, Technologies and Learning*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 525–536.
16. Krasadakis, P.; Sakkopoulos, E.; Verykios, V.S. A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics* **2024**, *13*, 648. [CrossRef]
17. Papantoniou, K.; Tzitzikas, Y. NLP for The Greek Language: A Longer Survey. *arXiv* **2024**, arXiv:2408.10962.
18. Fang, C.; Markuzon, N.; Patel, N.; Rueda, J.D. Natural language processing for automated classification of qualitative data from interviews of patients with cancer. *Value Health* **2022**, *25*, 1995–2002.
19. Cheng, K.; Inzer, S.; Leung, A.; Shen, X.; Perlmutter, M.; Lindstrom, M.; Chew, J.; Presner, T.; Needell, D. Multi-scale Hybridized Topic Modeling: A Pipeline for analyzing unstructured text datasets via Topic Modeling. *arXiv* **2022**, arXiv:2211.13496.
20. Liu, A.; Sun, M. From Voices to Validity: Leveraging Large Language Models (LLMs) for Textual Analysis of Policy Stakeholder Interviews. *arXiv* **2023**, arXiv:2312.01202.
21. Papaloukas, C.; Chalkidis, I.; Athinaios, K.; Pantazi, D.A.; Koubarakis, M. Multi-granular legal topic classification on greek legislation. *arXiv* **2021**, arXiv:2109.15298.
22. Mastrokostas, C.; Giarelis, N.; Karacapilidis, N. Social Media Topic Classification on Greek Reddit. *Information* **2024**, *15*, 521. [CrossRef]
23. Alexandridis, G.; Varlamis, I.; Korovesis, K.; Caridakis, G.; Tsantilas, P. A survey on sentiment analysis and opinion mining in greek social media. *Information* **2021**, *12*, 331. [CrossRef]
24. Pitenis, Z.; Zampieri, M.; Ranasinghe, T. Offensive language identification in Greek. *arXiv* **2020**, arXiv:2003.07459.
25. Michailidis, P.D. A Comparative Study of Sentiment Classification Models for Greek Reviews. *Big Data Cogn. Comput.* **2024**, *8*, 107. [CrossRef]
26. Loukas, L.; Stogiannidis, I.; Diamantopoulos, O.; Malakasiotis, P.; Vassos, S. Making llms worth every penny: Resource-limited text classification in banking. In Proceedings of the Fourth ACM International Conference on AI in Finance, New York, NY, USA, 27–29 November 2023; pp. 392–400.
27. Voukoutis, L.; Roussis, D.; Paraskevopoulos, G.; Sofianopoulos, S.; Prokopidis, P.; Papavasileiou, V.; Katsamanis, A.; Piperidis, S.; Katsouros, V. Meltemi: The first open large language model for greek. *arXiv* **2024**, arXiv:2407.20743.
28. Mylonas, N.; Stylianou, N.; Tsikrika, T.; Vrochidis, S.; Kompatsiaris, I. A Multi-Task Text Classification Pipeline with Natural Language Explanations: A User-Centric Evaluation in Sentiment Analysis and Offensive Language Identification in Greek Tweets. *arXiv* **2024**, arXiv:2410.10290.
29. Kydros, D.; Argyropoulou, M.; Vrana, V. A content and sentiment analysis of Greek tweets during the pandemic. *Sustainability* **2021**, *13*, 6150. [CrossRef]
30. Chalkidis, I.; Fergadiotis, M.; Androutsopoulos, I. MultiEURLEX—A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv* **2021**, arXiv:2109.00904.
31. Nikiforos, M.N.; Voutos, Y.; Drougani, A.; Mylonas, P.; Kermanidis, K.L. The modern Greek language on the social web: A survey of data sets and mining applications. *Data* **2021**, *6*, 52. [CrossRef]

32. Jimoyiannis, A.; Koukis, N. Exploring teachers' readiness and beliefs about emergency remote teaching in the midst of the COVID-19 pandemic. *Technol. Pedagog. Educ.* **2023**, *32*, 205–222.
33. Kostas, A.; Paraschou, V.; Spanos, D.; Sofos, A. Emergency Remote Teaching in K-12 Education During COVID-19 Pandemic: A Systematic Review of Empirical Research in Greece. In *Research on E-Learning and ICT in Education: Technological, Pedagogical, and Instructional Perspectives*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 235–260.
34. Lavidas, K.; Apostolou, Z.; Papadakis, S. Challenges and opportunities of mathematics in digital times: Preschool teachers' views. *Educ. Sci.* **2022**, *12*, 459. [CrossRef]
35. Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
36. Adoma, A.F.; Henry, N.M.; Chen, W. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In Proceedings of the 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 18–20 December 2020; pp. 117–121.
37. Liu, Y. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
38. Jojoa, M.; Eftekhar, P.; Nowrouzi-Kia, B.; Garcia-Zapirain, B. Natural language processing analysis applied to COVID-19 open-text opinions using a distilBERT model for sentiment categorization. *AI Soc.* **2024**, *39*, 883–890. [CrossRef] [PubMed]
39. Khan, L.; Amjad, A.; Ashraf, N.; Chang, H.T. Multi-class sentiment analysis of urdu text using multilingual BERT. *Sci. Rep.* **2022**, *12*, 5436. [CrossRef]
40. Anggrainingsih, R.; Hassan, G.M.; Datta, A. BERT based classification system for detecting rumours on Twitter. *arXiv* **2021**, arXiv:2109.02975.
41. Giarelis, N.; Mastrokostas, C.; Siachos, I.; Karacapilidis, N. A review of Greek NLP technologies for Chatbot development. In Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics, New York, NY, USA, 24–26 November 2023; pp. 15–20.
42. Koutsikakis, J.; Chalkidis, I.; Malakasiotis, P.; Androutsopoulos, I. Greek-bert: The greeks visiting sesame street. In Proceedings of the 11th Hellenic Conference on Artificial Intelligence, New York, NY, USA, 2–4 September 2020; pp. 110–117.
43. Wikipedia. Βικιπαίδεια: Αντίγραφα της βάσης δεδομένων. Available online: https://el.wikipedia.org/wiki/%CE%92%CE%B9%CE%BA%CE%B9%CF%80%CE%B1%CE%AF%CE%B4%CE%B5%CE%AF%CE%B1:%CE%91%CE%BD%CF%84%CE%AF%CE%B3%CF%81%CE%B1%CF%86%CE%B1_%CF%84%CE%B7%CF%82_%CE%B2%CE%AC%CF%83%CE%B7%CF%82_%CE%B4%CE%B5%CE%B4%CE%BF%CE%BC%CE%AD%CE%BD%CF%89%CE%BD (accessed on 14 January 2025).
44. Europarl. Europarl: A Parallel Corpus for Statistical Machine Translation. Available online: https://www.statmt.org/europarl/ (accessed on 14 January 2025).
45. OSCAR Project. OSCAR: Open Super-Large Crawled Aggregated Repository. Available online: https://oscar-project.org/ (accessed on 14 January 2025).
46. Athinaios, K.; Chalkidis, I.; Pantazi, D.A.; Papaloukas, C. Named Entity Recognition Using a Novel Linguistic Model for Greek Legal Corpora Based on BERT Model. 2020. Available online: https://pergamos.lib.uoa.gr/uoa/dl/object/2927727 (accessed on 14 January 2025).
47. Apostolopoulou, A.G.; Briakos, S.A.; Pantazi, D.A. Nlp Tasks with Greeklegalbert v2. Ph.D. Thesis, School of Sciences Department of Informatics and Telecommunications, Athens, Greece, 2021.
48. Kotsifakou, K.M.; Sotiropoulos, D.N. Greek political speech classification using BERT. In Proceedings of the 2023 14th International Conference on Information, Intelligence, Systems & Applications (IISA), Volos, Greece, 10–12 July 2023; pp. 1–7.
49. Evdaimon, I.; Abdine, H.; Xypolopoulos, C.; Outsios, S.; Vazirgiannis, M.; Stamou, G. Greekbart: The first pretrained greek sequence-to-sequence model. *arXiv* **2023**, arXiv:2304.00869.
50. Giarelis, N.; Mastrokostas, C.; Karacapilidis, N. GreekT5: Sequence-to-Sequence Models for Greek News Summarization. In *Proceedings of the IFIP International Conference on Artificial Intelligence Applications and Innovations*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 60–73.
51. Papadopoulos, D.; Metropoulou, K.; Papadakis, N.; Matsatsinis, N. FarFetched: Entity-centric Reasoning and Claim Validation for the Greek Language based on Textually Represented Environments. In Proceedings of the 12th Hellenic Conference on Artificial Intelligence, New York, NY, USA, 7–9 September 2022; pp. 1–10.
52. LightEternal. Lighteternal/nli-xlm-r-greek. 2023. Available online: https://huggingface.co/lighteternal/nli-xlm-r-greek (accessed on 13 January 2025).
53. Cao, J.; Xu, X.; Yin, X.; Pan, B. A risky large group emergency decision-making method based on topic sentiment analysis. *Expert Syst. Appl.* **2022**, *195*, 116527. [CrossRef]
54. Ahne, A.; Fagherazzi, G.; Tannier, X.; Czernichow, T.; Orchard, F. Improving diabetes-related biomedical literature exploration in the clinical decision-making process via interactive classification and topic discovery: Methodology development study. *J. Med. Internet Res.* **2022**, *24*, e27434. [CrossRef]
55. Huang, Y.; Wang, R.; Huang, B.; Wei, B.; Zheng, S.L.; Chen, M. Sentiment classification of crowdsourcing participants' reviews text based on LDA topic model. *IEEE Access* **2021**, *9*, 108131–108143. [CrossRef]

56. Tzimiris, S.; Nikiforos, M.N.; Nikiforos, S.; Kermanidis, K.L. Challenges and Opportunities of Emergency Remote Teaching: Linguistic Analysis on School Directors' Interviews. *Eur. J. Eng. Technol. Res.* **2023**, 53–60.. [CrossRef]
57. Kvale, S. *Interviews: Learning the Craft of Qualitative Research Interviewing*; Sage: Thousand Oaks, CA, USA, 2009.
58. Joffe, H. Thematic analysis. In *Qualitative Research Methods in Mental Health and Psychotherapy: A Guide for Students and Practitioners*; Wiley-Blackwell: Chichester, UK, 2011; pp. 209–223.
59. Grimmer, J.; Stewart, B.M. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Polit. Anal.* **2013**, *21*, 267–297. [CrossRef]
60. Commons, C. Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. 2025. Available online: https://creativecommons.org/licenses/by-nc-nd/4.0/ (accessed on 13 January 2025).
61. Tzimiris, S.; Nikiforos, S.; Kermanidis, K.L. Post-pandemic pedagogy: Emergency remote teaching impact on students with functional diversity. *Educ. Inf. Technol.* **2023**, *28*, 10285–10328.
62. Jimoyiannis, A.; Koukis, N.; Tsiotakis, P. Shifting to emergency remote teaching due to the COVID-19 pandemic: An investigation of Greek teachers' beliefs and experiences. In Proceedings of the Technology and Innovation in Learning, Teaching and Education: Second International Conference, TECH-EDU 2020, Vila Real, Portugal, 2–4 December 2020; Proceedings 2; Springer: Berlin/Heidelberg, Germany, 2021; pp. 320–329.
63. Zagkos, C.; Kyridis, A.; Kamarianos, I.; Dragouni, K.E.; Katsanou, A.; Kouroumichaki, E.; Papastergiou, N.; Stergianopoulos, E. Emergency remote teaching and learning in Greek universities during the COVID-19 pandemic: The attitudes of university students. *Eur. J. Interact. Multimed. Educ.* **2022**, *3*, e02207.
64. Beck, C.; Booth, H.; El-Assady, M.; Butt, M. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In Proceedings of the 14th Linguistic Annotation Workshop, Barcelona, Spain, 12 December 2020; pp. 60–73.
65. Jusoh, S. A study on NLP applications and ambiguity problems. *J. Theor. Appl. Inf. Technol.* **2018**, *96*, 345–350.
66. Haviana, S.F.C.; Mulyono, S.; Badie'Ah. The Effects of Stopwords, Stemming, and Lemmatization on Pre-trained Language Models for Text Classification: A Technical Study. In Proceedings of the 2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Palembang, Indonesia, 20–21 September 2023; pp. 521–527.
67. Nijhawan, L.P.; Janodia, M.D.; Muddukrishna, B.; Bhat, K.M.; Bairy, K.L.; Udupa, N.; Musmade, P.B. Informed consent: Issues and challenges. *J. Adv. Pharm. Technol. Res.* **2013**, *4*, 134–140.
68. Saunders, B.; Kitzinger, J.; Kitzinger, C. Anonymising interview data: Challenges and compromise in practice. *Qual. Res.* **2015**, *15*, 616–632.
69. HiLab, I.U. HiLab Datasets. 2025. Available online: https://hilab.di.ionio.gr/index.php/en/datasets/ (accessed on 13 January 2025).
70. Cateni, S.; Colla, V.; Vannucci, M. Improving the stability of the variable selection with small datasets in classification and regression tasks. *Neural Process. Lett.* **2023**, *55*, 5331–5356.
71. Nguyen, N.B.C.; Karunaratne, T. Learning Analytics with Small Datasets—State of the Art and Beyond. *Educ. Sci.* **2024**, *14*, 608. [CrossRef]
72. Nikiforos, S.; Anastasopoulou, E.; Pappa, A.; Tzanavaris, S.; Kermanidis, K.L. Motives and barriers in Emergency Remote Teaching: Insights from the Greek experience. *Discov. Educ.* **2024**, *3*, 1–21. [CrossRef]
73. Nikiforos, S.; Anastasopoulou, E.; Pappa, A.; Tzanavaris, S.; Kermanidis, K.L. Teachers' Needs for Support during Emergency Remote Teaching in Greek Schools: Role of Social Networks. *Computers* **2024**, *13*, 177. [CrossRef]
74. Zheng, Z.; Lu, X.Z.; Chen, K.Y.; Zhou, Y.C.; Lin, J.R. Pretrained domain-specific language model for natural language processing tasks in the AEC domain. *Comput. Ind.* **2022**, *142*, 103733.
75. Yu, S.; Su, J.; Luo, D. Improving bert-based text classification with auxiliary sentence and domain knowledge. *IEEE Access* **2019**, *7*, 176600–176612. [CrossRef]