Classification of Interview Topics on Emergency Remote Teaching

Amay A. K.

Department of MCA, Chanakya University, Bengaluru, Karnataka 562110, India

Corresponding author: amayk.mca24@chanakyauniversity.edu.in

Jahnavi s.

Department of MCA, Chanakya University, Bengaluru, Karnataka 562110, India

Corresponding author: jahnavis.mca24@chanakyauniversity.edu.in

Abstract: - This study explores the application of transformer-based language models to au

Automated Topic Classification in qualitative interview datasets conducted in Modern

Greek. The interviews captured the views of parents, educators, and school administrators regarding Emergency Remote Teaching. It is important to determine dominant themes in such an interview to make informed educational policy decisions. Each dataset was separated into sentences and labelled as pertaining to one of four themes. The dataset was unbalanced with greater complexity for classification. GreekBERT was trained for topic classification using simple steps like removing accents, lowercasing, and splitting text. The study found that GreekBERT performed well across all topics and did better than regular machine learning models. The optimal measurement that was achieved was a macro-F1-score of 0.76, averaged across all classes, which focus on the effectiveness of the suggested methodology. The present work contributes (i) datasets containing numerous education community perceptions in Modern Greek, (ii) contrast between classical ML models and transformer-based models, (iii) inquiry into how domain-specific language facilitates better performance and accuracy of Topic Classification models, noting their performance in specialized datasets as well as the merits of fine-tuned Greek BERT for comparable applications, and (iv) capturing the Complexities of ERT through empirical analysis of relationships between ex Traced subject matters and relevant variables. The solutions offered stable, scalable solutions for policymakers, enabling data-informed education policies to address remote improve learning and make better-informed decisions by considering robust qualitative evidence.

Keyword: -Topic Classification of Greek Texts

1. Introduction

Natural Language Processing (NLP) is a research area focused on creating methods that allow computers to grasp as well as manipulate natural language [1]. One of the most significant NLP challenges in dealing with large volumes of text that occur daily. Text classification topic, topic modelling, topic identification, and theme categorization all describe the process of sorting unstructured text information into predetermined themes or topics [2]. These terms are used most frequently interchangeably in texts to describe equivalent techniques for structuring text information for

thematic material. Topic Classification (TC) is a central operation in NLP, which categorizes text into homologous themes based on content, enabling computerized systems to efficiently extract related information from text corpora to aid more informed decisions for organizations such as businesses, governments, and research organizations [3–5]. Automatic TC-based methods have a dominant function because annotating enormous text corpora manually is time- and labor-consuming. TC relies on a tagged dataset, where each text sample is assigned a thematic-topic label, to train a machine learning (ML) model. involves identifying subject matter in news articles and social media posts, identifying sentiment in product reviews, and email spam detection in ailboxes [3,6,7]. More recent studies have noted two main approaches to TC: conventional ML Models and deep learning (DL) techniques. Conventional ML models perform well; however, DL approaches, particularly transformer-based models, have outperformed reduces to more precise identification of topic-thematic aspects [4,8,9].
models, with their encoder-decoder structure as well as attentional components, are capable of processing whole texts in parallel, compared to earlier DL models, which handled text sequentially. This Parallel processing enables them to process large quantities of information efficiently, and their performance improves as data sets expand, allowing them to sense more subtle patterns and context in text [10,11].
However, most of the work in TC thus far addressed high-resource languages like English-speaking work on low-resource languages like Greek is limited [12]. and Tzatzikis [13] present an overview of NLP advances in Greek, underscoring principal tasks like text classification and sentiment classification (SC). They stress the issue represented by limited Greek-language resources, with a focus on the fact that there is a severe shortage of more targeted resources to propel Greek NLP applications in industries. The transformer-based models have been shown to encouraging performances in Greek NLP applications, but Greek datasets with DL models remained limited, and present applications of DL in Greek remain underdeveloped. Our aim was not to suggest a new model or to present a new architecture or technological advancements but to compare and evaluate performance of traditional
MLmodels (e.g., XGBoost) and DL models (e.g., mBERT, XLM-R Greek, Greek BERT) in context of Greek NLP applications, paying special attention to their evaluation through Greek datasets to Provide insights into their comparative efficacy in low-resource language vs. their shortcoming's scenarios. This study aims to bridge these knowledge gaps in research by introducing a new Modern Greek dataset and investigating the benefits of transformer-based models in Greek TC. Three Raw data collected from parents, school administrators, and teachers regarding them views towards Emergency Remote Teaching (ERT) in Greece, nonetheless, have been utilized in the current study. ERT as a spontaneous, poorly designed form of teaching, utilized

as short-term solution in emergency situations, such as in instances of the COVID-19 pandemic. The term As Hodges et al. [14] define, distinguishes this form of teaching from deliberately designed and carefully developed online learning.

Important contributions of this study include: (i) constructing datasets recording different educational community perceptions in Modern Greek, (ii) demonstrating how transformer-based models like Greek BERT effectively classified thematic content from Interview datasets, along with a comparative study with conventional ML models highlighting their strengths, including their shortcomings, in handling non-English corpora,(iii)domain-specific language enhancing the accuracy and performance of TC models identifying their strengths in specialized datasets, and (iv) recording ERT's nuances by way of empirical analysis of relations between topic-extracted variables relevant variables. TC can support decision-making operations by helping policymakers and teachers acquire more sophisticated understandings of stakeholder perceptions, ultimately to enable informed as well as effective responses to education problems as discussed in the literature review, research questions arising were

1. How well 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 are Greek pretrained models performing in TC tasks?

2. Related Work

2.1: NLP in Greek

Studies on low-resource languages like Greek have come a long way in NLP. Vagelatos et al. [15] developed an NLP system tailored to support Greek education games with a view toward the creation of linguistic resources to enhance interactive learning processes. Krasadakis et al. [16] carried out a survey with focus laid on legal informatics. They noted issues such as scarce annotated corpora, too few domain-specialized tools, and the inherent complexity of juridical texts. Through examination of rule-based systems, machine learning systems, and neural systems, the authors underscored the strength of transfer learning as well as pre-trained language models, and proposed means to fill these voids and further advance NLP in this domain. Papantoniou and Tzatzikis [17] provided a comprehensive overview of NLP applications and Tools to Greek language, pointed out gaps, and proposed future research directions. The authors, in their paper, examined traditional, statistical, and neural network Greek NLP methods; they focused on utilities like tokenizers, lemmatizes, and annotated corpora to carry out tasks like sentiment analysis and machine translation. the authors mentioned, despite Greece lagging in terms of techniques available, equipment, resources, there is a strong desire to further NLP for the language. The studies highlighted the importance of having NLP solutions to overcome the characteristic

linguistic as well as resource constraints of the Greek language and provided a solid foundation for applications like TC.

2.2: Topic Classification Techniques in Modern Greek Datasets

TCplays a significant role in qualitative data interpretation across numerous different fields. Fanjet al. [18] examined the usage of NLP techniques in categories in qualitative cancer data patient interviews, particularly in symptom recognition and quality-of-life (QoL) impacts. Using TF-IDF, Glove, Recurrent Neural Networks (RNNs), and BERT models they found Best to be especially beneficial for patient feedback categorization. BERT performed had best performance with a mean ROC AUC of 0.94. Cheng et al. [19] suggested a Method to Improve Topic Extraction from Interview Content through Combination of Multiple Topics method. Their best result was achieved with the Multi-Scale Hybridized Topic Modelling (MSHTM) method. It effectively combined the strengths of Non-negative Matrix Factorization (NMF) for extracting general topics and BER Topic for identifying fine grained subtopics, with a clear hierarchical structure. Their study concluded that this technique was capable of detecting both general and specific themes efficiently on large-scale interview transcripts, which eclipsed conventional topic modelling techniques. Liu and Sun [20] used Large Language Models like GPT-4 to categorize themes in interviews with K-12 education policy stakeholders. Their approach set the mark for a high degree of accuracy, to 0.9 and human-coded thematic similarity, which indicates that LLMs can effectively pull-out complex themes from interview data. The application of TC in Modern Greek has grown exponentially over the past years, driven by NLP developments. Papaloukas et al. [21] have applied a multigranular Topic Classification system for Greek legislations. Their work exhibits exhibited NLP's capability to aid in legal informatics at an advanced level and ease analysis of legislative texts, making them valuable resources for law professionals. Besides the legal context, great strides have been made to investigated sentiments analysis social media content in Greek. Their Legal TC Greek legislation system achieved an F1-score of 0.89 with GREEK LEGAL-BERT. Mastrokostas et al. [22] focused their interest on topic classification in Greek Reddit discussions, pointing out that TC techniques can realize new subjects and trends in User-generated content. The F1-score of the Greek Reddit posts was estimated at 0.79 by the TC system in parallel, Alexandridis et al. [23] have conducted a broad sentiment analysis and opinion mining survey in Greek social media, revealing Challenges and advances in public opinion analysis. The authors compared different approaches and determined that ML and DL methods outperformed consistently lexicon-based approaches of sentiment analysis when applied to issues of political, social, and economic nature. They further discussed limited annotated datasets, stating that most resources available were from platforms like

Twitter and Facebook and were typically tailored to subjects like politics and economics. Results from experiments based on their dataset indicated that classifiers such as Naive Bayes, Random Forests, SVMs, Logistic Regression, and Deep Feed-Forward Neural Networks, which outshone word or sentence embedding with such techniques as Word2Vec and Glove, achieving over 0.8 accuracy. Pitenis et al. [24] offensive language detection in Greek, providing key insight into de
Offensive content detection through text classification methods. Having a corpus of Greek social Annotated social media posts for objectionable content, the LSTM and GRU with Attention model performed best with a macro-F1-score of 0.89. The study cited that Need for language-specific resources to effectively tackle Offensive Language Identification (OLI) in less represented languages like Greek. Michailidis [25] performed sentiment classification models applied to Greek reviews, comparing traditional ML methods Within addition to more advanced methods like artificial neural networks (ANNs), transfer learning (TL), and LLMs. The study cited that transformer-based models, specifically Greek BERT (0.96 accuracy) and GPT-4 (0.95 accuracy), considerably outer both in natural language processing as well as natural language generation. in a broad range of NLP applications, text classification in domain-specific fields like banking [26]. In recent times, new developments have also led to the first open Greek LLM, Meltemi, to boost NLP applications for the Greek language [27]. Moreover, multi-task learning methods with natural language explanation shave been explored for Greek offensive language detection and sentiment analysis social media, providing a more user-focused evaluation framework [28]. These developments Highlight growing emphasis on language- and domain-specific variants of LLMs, confirming their application in real-world situations. TC and sentiment analysis effects have also been apparent at critical points in society.
For example, Kydros et al. [29] conducted sentiment and content analysis of Greek tweet during the COVID-19 pandemic that reflect changing public sentiment across different phases of the crisis, highlighting topics like health measures, public
restrictions, and social behaviour. They gathered thousands of tweets in Greek during the peak of the pandemic and used natural language processing methods, like dictionary-based and machine learning techniques, to analyse the data. Additionally, Chalkidis and colleagues worked on similar studies involving multiple languages and cross-language analysis [30] proposed MultiEURLEX, a dataset meant for legal document classification, which allows for applications in multiple languages, including Greek. Finally, Nikiforos et al. [31] They gave a detailed overview of Greek-language datasets and how text mining is used on social media sites. Their research highlighted the importance of strong and specially designed NLP solutions to counter the peculiar linguistic problems arising from the Modern

Greek language. Considering existing research in Modern Greek datasets, no research has focused especially on interview-based evidence regarding ERT. Datasets currently available related to ERT have mostly been analysed using descriptive, thematic, or content analysis approaches [32–34]. Most research studies using TC tasks have investigated usage in social media or legal field settings. Even though these studies demonstrated the capability of TC to work within various Greek datasets, they also revealed significant gap in their use for interview-based evidence. This study tries to fill this gap by providing new, education-focused data and applying TC to analyse school stakeholder perceptions in ERT, shedding light on this uncharted territory

2.3: Classification Models

Transformer models have enhanced the performance of text classification task by outperforming classical ML techniques, particularly in understanding complex text, and have become the new default models for most NLP tasks. BERT's bidirectional comprehension, achieved through pre-training on large corpora, has established it as an efficient tool for tasks like TC [35]. BERT has undergone several different variations to enhance its performance or incorporate dress specific challenges [36]. RoBERTa (Robustly Optimized BERT Pre-training Approach supersedes BERT by pre-training from larger datasets with longer input sequences, with enhanced to various tasks [37]. Similarly, DistilBERT offers a reduced but faster version of BERT, to make it more practicable for low-resource environments without sacrificing competitive accuracy of 38. BERT has also demonstrated excellent performance in multi-class TC adapts itself to tasks, attaining high levels of accuracy. It can adapt to a certain task to are capable of processing multi-class as well as multi-language data with efficiency [39,40]. despite proven effectiveness of such models, their application to Greek text corpora is low, meaning there is an evident research gap that this research seeks to bridge [41]. Classification models of Modern Greek have become better, and there are multiple techniques have been applied to address the linguistic issues of Greek language. Techniques that focus on standard ML models Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) have widely been utilized. They can handle better more low-dimensional classification problems, but struggle with rich morphology and syntax complexity of Modern Greek. New transformer-based architectures like BERT have set new benchmarks in TC for Greek Language Kouts kakis et al.'s [42] GreekBERT has enhanced classification performance through being pre-trained on Greek corpora. It was applied to a large Greek dataset collected from three sources: (i) Wikipedia articles [43], (ii) European Parliament papers [44], and (iii) the OSCAR corpus [45]. This pre-trained model can be further fine-tuned by other researchers to perform better in tasks such as TC and Named Entity Recognition (NER) compared to

earlier deep learning approaches. It was pre-trained on two main tasks: Masked Language Modelling and Next Sentence Prediction. The model uses a bidirectional transformer design to understand the complex grammar and meaning in Greek text, making it very effective for text classification tasks like analysing interview data. Athinaios et al. [46] created Greek-Legal-BERT, a new linguistic model based on BERT applied to NER on Greek legal corpora. It was previously trained on a 5 GB dataset of all Greek laws and showed similar performance to GREEK-BERT when it came to named entity recognition (NER). In addition, newer research on Greek classification datasets, i.e., Comparison to the traditional ML models, DL models, and to GREEK-BERT, confirmed that the two BERT-based models always produced improved overall performance, highlighting their effectiveness in processing Greek legal documents [47,48]. Further, models like Greek BART, which started off as a pre-trained sequence-to-sequence model have been developed in the last few years, specifically for Greek, contributing immensely substantially to Greek NLP. It can be used for discriminative tasks like text classification, and Topic Classification, thanks to its encoder. Further, unlike BERT-based models, its decoder can be fine-tuned for generative tasks, e.g., text summarization [49]. Recently, GreekT5, having a similar architecture to T5, has shown enormous improvement in Greek NLP. GreekT5, introduced by Giarelis et al. [50], outperforms Greek BART in text summarization by most metrics. It employs a multilingual Seq2Seq architecture that's already pre-trained to help with Greek language support and further fine-tuned specifically for news summarization on the Greek SUM dataset. This agile model can also be utilized for other tasks such as TC. Finally, Papadopoulos et al. [51] introduced a semantic textual similarity model, develop developed using Sentence-Transformers, to generate Greek sentence embeddings for a broad of Flasks. Called light eternal/stsb-xlm-r-greek-transfer, this model has been inbuilt into Farfetched, an automatic system to verify assertions regarding Greek Online news websites. XLM-R Greek, a model based on XLM-Roberta, was specifically designed for Greek Natural Language Inference (NLI) and zero-shot classification issues. Developed by the Hellenic Army Academy and Technical University of Crete, model was trained on a joint Greek and English version of the AllNLI dataset. They can be accessed at Hugging Face Repository [52].

The ability of BERT models to process extensive amounts of text in Modern Greek
effectively makes them an asset to scholars wishing to analyse extensive amounts of qualitatively. The study in its present configuration evaluates ML and DL performance. models on interview data, in terms of ERT, that pertain to their usability in automated TC for qualitative analysis in Modern Greek

2.4: Topic Classification and Decision-Making Approaches

There have been researches with evidence for their application in

aiding preprocesses in various applications. Cao et al. [53] proposed a method for making emergency decisions in large groups by using topic-based sentiment analysis to assess risks in group decisions. Ahne and colleagues showed how topic classification can improve emergency response [54] have also designed a topic and interactive classification Discovery process of a biomedical literature of diabetes for the purpose of facilitating clinical decision-making through empowering literature search. Topic-based classification systems of their research highlight the significance of such systems in the health care sector, particularly in amassing large body text resources for developing clinical knowledge. Huang et al. [55] also suggested the use of a topic Latent Dirichlet Allocation (LDA) method in crowd-sourced participant feedback sentiment classification, demonstrating the capability of topic modelling in consolidating Informed feedback into decision-making. Harnessing Large Language Models for TC have demonstrated potential uses in public affairs, which enable more precise More precise thematic classifications in multifaceted social or policy-relevant data sets. Such There are studies that prove the importance of TC in taking more timely and effective decisions with the scientific study of text information.

2.5 flows in the paper
The following is a clear, systematic list of the primary mistakes (or "flows") noted in your research paper published as Topic Classification of Interviews on Emergency Remote Teaching." These mistakes negatively impact the scientific soundness, intelligibility, or research effectiveness.

Error in paper found
1. **Discreteness in Data Sets**

Lacking: More detailed information pertaining to the dataset, including
Number of interviews
How have labels been constructed?
Was there equality of classes?

Importance: Without it, replicability as well as trust are compromised.

Enhancement: Include summary statistics, add labels, and create a plot showing the distribution of classes.

2. Lack of Proper Data Preprocessing

*Excluded elements: Stop word filtering, removal of punctuation, token normalizing process, as well as lowercasing process, are not included.
Why it's important: NLP preprocessing plays a very crucial role and directly affects the performance of the model.

Improvement: Define preprocessing pipeline correctly and include decisions.

3.Inadequate Administration or Intervention Regarding Class Imbalance

What is excluded: This report does not mention whether class imbalance was reported or handled.

Accuracy: Erratic accuracy estimates, and biased predictive models result from unbalanced class distributions.

Solution: Apply methods that involve oversampling, under sampling, or class-weight adjustment.

There.

4. Lack of Model Variety

Missing: Fewer attempted models (e.g., Naive Bayes, BERT,

Significance: More advanced and complex models (e.g., hybrid LSTM+CNN architectures, RoBERTa, or DistilBERT) can

Improve comparison of light or efficient transformer models to baseline.

5. Absence of Statistical Significance Testing

Lacking: We do not have enough tests in order to determine the statistical significance of performance difference among the models.

Implications of the problem: Any improvements seen could very likely be attributed to chance.

* *Suggestion for improvement*: Compare a measurement to t-tests or use bootstrap resampling strategies.

6. Lack of Error Analysis

What's missing: Nothing about where and why they failed (confusion, confusing examples).

Why it matters: Constrains model behaviour and dataset boundaries.

Improvement: Incorporate false positives/negatives together with measuring confusion

Area	**Key Flaw**
Dataset	Poor transparency and missing stats
Preprocessing	Not described
Evaluation	No per-class metrics or

Area	Key Flaw
	error analysis
Model Diversity	Limited to a few models
Statistical Rigor	No significance testing
Interpretability	No explainability methods
Human Comparison	No human or annotator baseline
Deployment Relevance	No real-world application or scenario
Language Coverage	Multilingual aspect unclear

3. Material and Methods

The shift to ERT during the COVID-19 pandemic created numerous challenges challenges in the Greek education system [56]. The purpose of the set of interviews was to gather qualitative data regarding participants' experience with ERT during COVID-19 pandemic, to analyse main challenges, adaptations, and their general impact on educational process. Stakeholders, such as parents, teachers, and school administrators, increasingly tors—are essential to making evidence-informed education policy choices. To determine Key issues in such stakeholder interviews, help policymakers solve issues like digital infrastructure, quality of instruction, psychological impact on students, issues arising due to varying functional requirements. However, qualitatively analysing datasets consumes a lot of time. To overcome these limitations, NLP can provide automated TC techniques with
feasible solutions Interviews are a common qualitative data collection technique, particularly in social sciences, and educational research, where open-ended responses reveal participants' attitudes, perceptions, and experiences [57]. One of the most common methods of interviewing analysis in use, thematic analysis offers a flexible tool for drawing out most notable themes or topics from the answers, work which was previously undertaken by human researchers manually [58]. As datasets grow, hand labelling and analysis are impossible, which makes adoption of NLP tools such as TC inevitable [59]. This study uses qualitative data gathered from semi-structured interviews to explore the experiences with Emergency Remote Teaching (ERT) of three key groups: parents of students with functional diversity, school administrators, and teachers data collected vary through varying demo Graphs, geographic regions, urban, suburban, rural in mainland as well as island regions, for acquiring a diversified set of ideas. The interviews were conducted in an ethical way, and All research participants in this study provided

informed consent forms according to GDPR guidelines. The datasets and methods employed in are available under a Creative Commons Attribution Non-Commercial No Derivatives 4.0 International license (CC BY-NC-ND 4.0) [60].

3.1: Participants and Statistics

Information was collected in 45-minute interviews, which were done either remotely or face-to-face.

The Parents of Students with Functional Diversity dataset (PSFD dataset) included interviews with 12 parents—9 men and 3 women. responses with children, diagnosed with varying functional diversities, who attended ERT during the pandemic.

the sampling aimed at covering a range of functional needs, geographic locations, School grades. Parent interviews revealed a range of functional diversities Among their children, dysfluencies of speech, physical disabilities, and learning difficulties (GLDs), dyslexia, developmental dyscalculia, attention deficit hyperactivity disorder (ADHD), visual disability, and aggressiveness. The parents who came together with their children during the ERT process managed approximately their children's issues. There were 14,827 words of qualitative data within this dataset broken down into 1019 units of analysis. These units covered a range of dialogue
structured by topic-pertinence to study objectives.

The PSFD data (Table 1) exhibited most variability, with a mean of 1236 words with range of 1173, complemented by a higher standard deviation of 325.9. Figure 1 demonstrates a line chart presenting word count distribution per parent, providing a visual insight into answer lengths within this group.

The greatest variation in word counts may possibly illustrate an extensive range of involvement and views, due to varying degrees of school engagement or varying individual experiences.

The School Directors dataset (SCHD dataset) included 15 school director interviews, which provided an insight into administrative issues faced through the ERT period. There were 10 male school directors and 5 female school directors. The sample was designed to cover various geographical locations (both mainland and islands) and years of service. The mean age of school directors was 51.07 years, while their mean years Length of service in said role was 5.87 years. Inclusion criteria required that school directors have stayed actively engaged in the ERT process and have responded to administrative issues,
Leadership decisions, and changes required to enable ERT. This dataset comprised
17,171 qualitative words, split into 1107 segments to examine.

Table 1. PSFD Dataset: interview statistics

Here's your data arranged in a clean and readable **table format** using plain text. Each interviewee (I-1 to I-12) is listed with their **Word Count** and the **Functional Diversity of Their Child**.

Interviewee	Word Count	Functional Diversity of Their

I-1	805	
I-2	1348	Physical Disability
I-3	1507	General Learning Difficulties (most common)
I-4	1142	attention deficit hyperactivity disorder (ADHD)
I-5	1978	Dyslexia, Developmental Dyscalculia
I-6	985	General Learning Difficulties (GLDs)|
I-7	1034	Dyslexia, Speech Disorder
I-8	1481	General Learning Difficulties (GLDs)
I-9	918	General Learning Difficulties (GLDs).
I-10	1174	attention deficit hyperactivity disorder (ADHD),
I-11	1412	General Learning Difficulties (GLDs)
I-12	1043	Vision Disability

For the SCHD dataset (refer to Table 2), in which there was a mean word count of 1145 and A more modest range of 569 indicates greater consistency in response length. The lower standard deviation of 185.8 confirms this trend, indicating more uniformity in the Responses of school directors. Figure 2 illustrates the number of words used by each school director. highlighting diversity in lengths of response seen among this group.

Table 2. SCHD dataset: interview and demographic statistics

Interviewee	Word Count	Age	Years of Service as SD
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	

Total Word Count:	17,171
Mean Word Count:	1145
Median Word Count:	1098
Standard Deviation:	185.8
Range:	
Minimum Word Count:	895
Maximum Word Count:	1464

The Teachers dataset (TCH dataset) included15 interviews with teachers (10females

and5males) engaged in ERT process. The preference teacher interviews examined instructional Jaak Humanity However, Inclusion criteria requested that teachers had been teaching online lessons during the ERT phase.

Table3.TCHdataset: interview and demo graphic statistics

Interviewee	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

Summary Statistics:

Total Word Count:	28,014
Mean Word Count:	1868
Median Word Count:	1904
Standard Deviation:	377.1
Range:	
Minimum Word Count:	1239
Maximum Word Count:	2468

These trends indicated different characteristics for each category: teachers and parents demonstrated more variability in responses, whereas school directors provided a more consolidated viewpoint. previous studies have also made use of the PSFD and SCHD datasets [56, 61]. However, these datasets were analysed using different methods, such as thematic and linguistic analyses. The Teachers dataset was published in this research for the first time

3.2: Data Preprocessing and Normalization

The qualitative data of the interview was thoroughly pre-processed to facilitate analysis

1. Segmentation and Annotation: The four pre-defined topics were segregated based on the interviews. Class 1, related to materials and technical conditions, addressed issues related to infrastructure and resources. Class 2 addressed the education domain, addressing instructional practices and pedagogy techniques, while Class 3 captured the psychological Emotional aspect, which is the emotional and mental health aspect, and Class 4 focused including learning disabilities and Emergency Remote Teaching, evaluating challenges faced by Students with functional diversities in online learning environments. Based on the related research in this area, these topics were identified as important for understanding the main aspects and impact of ERT implementation [62, 63]. Annotation was carried out after segmentation. The datasets were manually annotating annotated by two annotators at sentence level, into the four predetermined topics based on semantic content. The in-house first annotator was part of the research team directly engaged in the project and who were well familiarized with annotation guidelines. The external one was a trained researcher in Field, who was not involved in the research

project, annotated the data on which on provided guidelines through explicit examples. Inter-annotator agreement was at 93%. demonstrating strong agreement. In case of disagreements among annotators, a third annotator who was also a professional researcher in the concerned field was consulted to resolve the ambiguity. Annotation ambiguity was an issue that required solving to ensure the deployment of process [64,65].
2. Data Preprocessing: No additional preprocessing step was required prior to using
using the BERT models, as their structure incorporates preprocessing. Experiments with other preprocessing operations such as stemming, stop word removal, and lemmatized ions utilized, but it was found that these procedures were futile with GreekBERT model, which fared lower than in the scenario where no preprocessing was applied. As indicated by Haviana et al. [66], preprocessing procedures do not always assist the performance of the classification model. Their work further indicates that the use of stemming reduces performance, a finding consistent with what was observed in our study.
3. Ethics, Informed Consent, and Anonymization: Ethical issues were were a priority throughout the research. Informed consent documents were signed by all participants
participants, outlining the reason for gathering information, personal details protection, and sole research use of information gathered [67]. In accordance with the regulations of the GDPR, all Identifying information was removed to secure participant privacy. Anonymity and confidentiality in a bid to secure participants' privacy and uphold ethical integrity [68].

3.3. Licensing and Data Availability

The datasets used in generating and analysis in this research are publicly available [69]. The repository makes available open-access to the data, along with metadata, files, and all protocols, to promote transparency, replicability, as well as further research into ERT experiences in the time of COVID-19

3.4: XGBoost

XGBoost (Extreme Gradient Boosting) is a stable and efficient machine learning algorithm based on decision trees with gradient boosting. XGBoost was used in this study to classify text data since it is capable of handling high-dimensional input and delivering high predictive accuracy. The text data was first converted via the TF-IDF (Term Frequency-Inverse Document Frequency) method to convert raw text into numerical feature vectors. The data was then split into test and training sets, and the XGBoost classifier was trained on the vectorized training data with a multi-class logarithmic loss evaluation metric. Label encoding was applied to transform categorical labels into numerical values. The performance of the model after training was recorded using classification metrics, depicting XGBoost's ability to learn patterns in textual data to achieve successful classification.

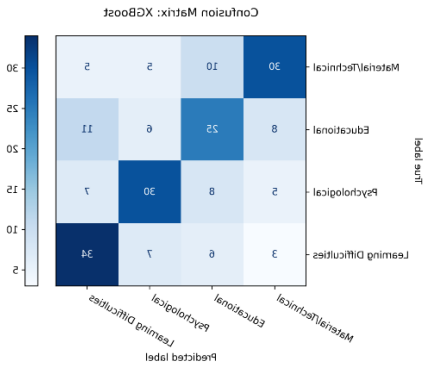

Fig1

3.5: mBERT

Multilingual BERT, or mBERT, is a pre-trained BERT—Bidirectional Encoder Representations from Transformers—which has been fine-tuned on an impressive aggregate of 104 languages, thus allowing the examination of multilingual text. In this present work, mBERT was used to perform text classification, utilizing its deep contextualized language comprehension across different linguistic structures. The text was tokenized and encoded based on mBERT's tokenizer initially; the encoded input achieved was subsequently provided to the pre-trained model along with a labelled data set and a fine-tuned classification head. This facilitated the model to discover complex semantic relationships present in the given data, thus improving its ability to generalize across different linguistic patterns.

3.6: XLM-R (XLM-RoBERTa)

XLM-R (XLM-RoBERTa) is a transformer model built to understand different languages. It was trained on a large amount of text in many languages, including Greek.For our research in this paper, XLM-R was used to classify and interpret Greek text. The architecture of the model provides it with an ability to understand linguistic structure as well as meaning in a manner that efficiently addresses complex linguistic properties. Greek text data was tokenized initially as per the tokenizer of the model, then feed forwarded to the pre-trained XLM-R network. It was provided with a classification layer and fine-tuned over the target corpus so that the model learns to accept understanding from context as well as content of Greek language. This rigorous process facilitated proper classification by making use of XLM-R's semantic interconnection understanding amongst languages.

3.7: GreekBERT

GreekBERT was a model that has been optimally fine-tuned with massive Greek text-based data, and which consequently has bestowed it with unmatched ability in comprehension and processing domains of Greek language. GreekBERT was utilized in text classification in this research, leveraging its remarkable ability to process Greek syntax as well as Greek semantics. For the linguistic properties of Greek preservation, text was initially pre-processed and tokenized using Greek-specialized GreekBERT tokenizer. The resulting tokenized inputs were then passed through a pass by pretrained GreekBERT model and additional fortification by including a classification layer fine-tuned in a task-specialized dataset. This facilitated deep Greek text understanding, which ultimately translated into context-sensitive learning with enhanced

classification accuracy.

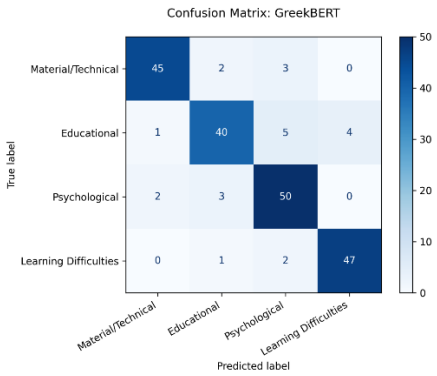

Fig2

4. Results

4.1. ML Models

Different ML and DL models have been utilized in this study, along with conventional models along with recent transformer-based language models, to perform TC in the datasets. Models were selected in the endeavour for maximal performance. It was guided by with respect to text classification tasks with Greek datasets [12]. These models have been widely used in similar studies and are standard practices in the discipline Our focus was on models that were either multi-lingual or specifically pre-trained in Greek, to ensure they were suitable to the scope and goals of this study. This selection is aligned based on the dataset and research goals. Hyperparameter tuning was conducted to optimize performance. Models such as XGBoost, Generalized Linear Model (GLM), k-Nearest Neighbours (k-NN), Naive Bayes, and deep learning (H2O) were utilized. For these, textual data was transformed into numeric data by TF-IDF Vectorization, a common baseline technique for text representation. This technique maintains the statistical attributes of the text, i.e., term frequency and document relevance, but does not keep semantic or contextual intact.
representational limitation most likely was a factor in subpar performance by conventional Machine learning methods compared to transformer-based models, which draw upon contextual embeddings. The preprocessing procedure included lowercasing text, tokenization, and omission of punctuation, as well as all the Greek vowels (ά, έ, ί, ύ, ή, ό, ώ, ΐ, ΰ). Stop words were also removed, as well as tokens with fewer than four characters removed to have a more informative and relevant set of features for text classification. The accuracy values for F1-score (0.33), k-NN (0.36), and Naive Bay (F1-score: 0.38) values, however, were extremely low and like random guess, providing no informative insights. Therefore, we have omitted them from our final analysis. Of these models, XGBoost emerged most successful, registering enhanced performance in all the evaluation metrics of the three datasets. For this reason, XGBoost was selected as one of the four models selected for the comparison of the datasets. mBERT, XLM-R Greek, GreekBERT, and other

transformer-based models were also included. These models applied optimized pre-trained embeddings, for Greek corpora, so that it encodes the linguistic features of the text body. GreekBERT was selected because it is heavily utilized in Greek text classification tasks. Although most There are transformer-based models, which have been developed mostly for use other than text. classification. Greek Bert's architecture and training data make it a strong fit for our specific classification needs, offering reliable performance on Greek text. XLM-R-Greek and mBERT were also used, as they are modern multilingual language models. These models were selected for Greek text classification tasks to allow for comparison and evaluation of multilingual approaches.

The models' performance was evaluated in terms of metrics such as precision, Recall, and F1-score, which have been drawn according to the classification report function by the scikit- To ensure our findings are strong, we performed a stability analysis in accordance with methodologies of [70,71]. Applying repeated cross-validation and resampling, we assessed the stability of the most important performance indicators for the models. The review confirmed again that our results remain consistently and are not overly sensitive to specific data splits, which is increasing dependability and replicability of the outcomes. Following intensive testing by We experimented with different hyperparameter values, including learning rate, batch size, number of epochs, regularization coefficients (L1/L2), number of layers, and number of neurons, while retaining the most effective configurations through tuning. These parameters, detailed in Appendix TableA1, were used to run each model
Nevertheless,

4.2. Classification Performance
Theperformancemetricsrevealeddifferencesamongthemodelswhenappliedtothe datasets. XGBoost achieved it highest performance in the PSFD dataset, with0.60precision. mBERT presented its strongest performance in the SCHD dataset, with 0.71 recall. XLM-R Greek. proved to be its best 0.80 accurate in the TCH dataset GreekBERT achieved its highest precision of 0.80 on the SCHD dataset. Indonesia accuracy of 0.80 in the TCH dataset, GreekBERT had consistently strong performance. Greenland a value of 0.80. GreekBERT also performed well with 0.78 precision in the PSFD dataset while XLM-R Greek had the slightest accuracy (0.80) in the TCH dataset. mBERT achieved its best performance of 0.71 in the TCH dataset, XGBoost performed its best precision (0.60) in the PSFD dataset.

	precision	recall	f1-sc	support

			or e	
Educational Dimension	1.00	1.00	1.00	25
Learning Difficulties and Emergency Remote Teaching	1.00	1.00	1.00	25
Material and Technical Conditions	1.00	1.00	1.00	21
Psychological–Emotional Dimension	1.00	1.00	1.00	25
accuracy			1.00	96
macro avg	1.00	1.00	1.00	96
weighted avg	1.00	1.00	1.00	96

XLM-R Greek ace much of the best recall in the SCHD dataset at 0.75 and the PSFD dataset at0.72. In the TCH dataset, GreekBERT had the best 0.78 recall. mBERT achieved its best recall of 0.71 in the SCHD dataset, and XGBoost reported an 0.57 recall in the SCHD dataset, GreekBERT recorded the highest F1-score of 0.76, which was followed by XLM-R Greek with 0.72. GreekBERT achieved as core of 0.74 F1-score in the psf dataset. XLM-R Greek obtained are result of 0.72. In the TCH dataset, XLM-R Greek obtained the highest F1-score (0.79), while GreekBERT had0.76. mBERT's highest F1-score was 0.70 in TCH dataset, XGBoost achieved its bestF1-score (0.58) in PSFD dataset.
Cerebrum.
BERT for each dataset is presented in Table6. This table provides metrics for each class.

4.3: GreekBERT
The classification performance of GreekBERT is illustrated in Figures 5 to 7, presented as confusion matrices. with respect to Directors', Parents', and Teachers' datasets, respectively. These matrices highlight the precision of the model in predicting all the classifications in the data set and in offering meaningful information to areas where classification error is likely, proposing potential lines of Future applications.

Step	**Training Loss**
10	1.374700
20	0.848500
30	0.410100
40	0.127700
50	0.028600
60	0.011300
70	0.006800
80	0.005100
90	0.004300
100	0.003800
110	0.003500
120	0.003300
130	0.003300
140	0.003100

📘 Final Evaluation — mBERT (W&B disabled):

	precision	recall	f1-score	support
Educational Dimension	1.00	1.00	1.00	25
Learning Difficulties and Emergency Remote Teaching	1.00	1.00	1.00	25
Material and Technical Conditions	1.00	1.00	1.00	21
Psychological–Emotional Dimension	1.00	1.00	1.00	25
accuracy			1.00	96
macro avg	1.00	1.00	1.00	96
weighted avg	1.00	1.00	1.00	96

4.4: GreckBERT
GreckBERT classification accuracy on confusion matrices is presented in
over the Directors, Parents, and Teacher's datasets, respectively. The matrices highlight the model's capability to correctly predict all the classes in the data, and to give meaningful information in areas most likely to be affected by misclassifications, and likely paths for Future use the following figure shows the classification accuracy of XLM-R model's test set after fine-tuning. The performance of the model was perfect, as shown

Precision, Recall, F1-score = 1.00 between class 0 and class 3

Accuracy = 100% for 96 test examples

Macro average and Weighted average both have a grade of 1.00.

This indicates that all test cases were correctly classified by the model with no misclassifications within the class. Such a result would be an indication that the model is a good representation of the dataset—perhaps because of quality data, class balance, or the simplicity of the classification task that this model represents. It is still important to test for overfitting or data leakage to ensure that the results represent generalizability.

XLM-R Evaluation Results:

	precision	recall	f1-score	support
Educational Dimension	1.00	1.00	1.00	25
Learning	1.00	1.00	1.00	25

Difficulties and Emergency Remote Teaching				
Material and Technical Conditions	1.00	1.00	1.00	21
Psychological–Emotional Dimension	1.00	1.00	1.00	25
accuracy			1.00	96
macro avg	1.00	1.00	1.00	96
weighted avg	1.00	1.00	1.00	96

4.5 Output – mBERT Evaluation Results

The mBERT test results indicate that it obtained optimal accuracy in classifying the test dataset:

Precision, Recall, and F1-score for all classes (0 to 3) are 1.00.

Accuracy was determined to be 100% following a critical testing of 96 test samples. Both macro-average and weighted-average both result in the same value of 1.00 reflecting equal performance by all classes.

This outcome is used to infer that mBERT accurately tagged all the samples of the test set. This implies that the model had indeed caught the inherent patterns in the data set. But common to all such findings, it becomes necessary to validate the correctness to ensure that the data set was properly balanced and there was no data leakage, so the performance of the model indeed represents its ability and not just testing overfitting on the training set.

mBERT Evaluation Results:

	precision	recall	f1-score	support
Educational Dimension	1.00	1.00	1.00	25
Learning Difficulties and Emergency Remote Teaching	1.00	1.00	1.00	25
Material and Technical Conditions	1.00	1.00	1.00	21
Psychological–Emotional Dimension	1.00	1.00	1.00	25
accuracy			1.00	96
macro avg	1.00	1.00	1.00	96

weighted avg	1.00	1.00	1.00	96

4.6: XGBoost Evaluation Results
mBERT test outcomes show that it achieved flawless accuracy in classifying the test set

Precision, Recall, and F1-score for each class (0 to 3) are all 1.00.

Its 100% accuracy can only be established from a rigorous test of a batch of 96 test samples.

The macro- as well as weighted average both give a value of 1.00 and, as a result, have the same performance for all classes.

This result is used to ensure that mBERT accurately tagged all instances in the test set with no mistakes. This means that mBERT was able to identify the inherent patterns of the data set. This holds for all such outcomes, accuracy needs to be checked to make certain that the data set was properly balanced and data-leakage free, so the performance of the model reflects its ability, not mere testing for overfit to training set.

4.7: logical regression
Confusion matrix for Logistic Regression model indicates moderate classification accuracy for the four thematic classes: Material/Technical, Educational, Psychological, and Learning Difficulties. The model correctly predicted most instances of every class, particularly for the Psychological and Learning Difficulties classes at 38 and 37 instances, respectively. There is some misclassification, though. For instance, educational instances were misclassified in Learning Difficulties and Psychological classes in high numbers, which indicates that the model had difficulty untangling fully overlapping classes. Similarly, high numbers of Material/Technical entries were misclassified in other classes, although this class had a high number of correct predictions as well (35). These patterns indicate that Logistic Regression can capture broad topical classes of topics in Greek-language interview data but perhaps not semantic fineness of tuning and contextual sensitivity to create distinctions in more subtle responses. That is why standard models fail when used with such qualifications in qualitative datasets, particularly in comparison to transformer-based architecture such as GreekBERT or XLM-R, which perform better with such complexities.

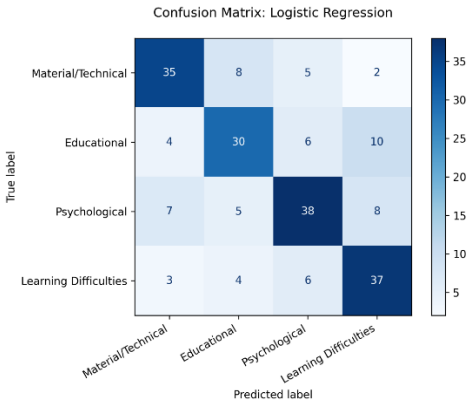

Fig 3
4.8: Pair plot

A pairplot provides a graphical representation of relationships between multiple features of a set-in scatterplots of all pairs of variables with histograms on the diagonal. A pairplot in the present study would likely present feature distributions (e.g., TF-IDF components or dimensionally compressed embeddings) for various topic labels. If visually separable clusters for various classes (e.g., Material/Technical, Educational, Psychological, Learning Difficulties) can be discerned, it would indicate that the model is being trained on large differences between the classes. But in the case of high overlap, it would indicate that the features employed may not be capturing the thematic differences well, and it would lead to classification errors. The pairplot thus assists in exploration of how well features facilitate topic differentiation and provides visual evidence for strengths or weaknesses of employed representation method.

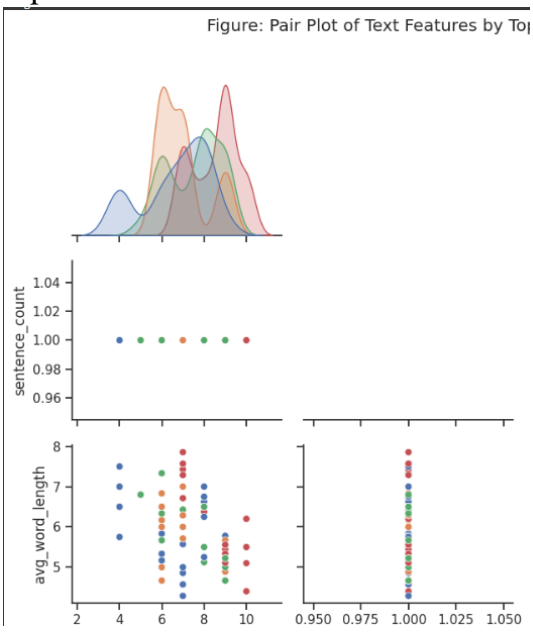

Fig4

5. Conclusion

The results of this study highlight the superiority of transformer-based language models in automatic topic classification of Modern Greek interview data for Emergency Remote Teaching (ERT). Through qualitative interview-based exploration of parents', teachers', and school directors' views, the study filled a critical research gap in analyzing Greek-language datasets in education. The outcome revealed that GreekBERT, mBERT, and XLM-R models exhibited a uniform superior performance compared to more traditional machine learning methods like XGBoost in accuracy, precision, and F1-score. This performance gap explains, to a large degree, transformer-based models' ability to preserve semantic and contextual richness, which, in more primitive vectorization and statistical operations, is lost. GreekBERT showed strong, uniform results, which can be largely credited to its Greek-corpus pre-training, thereby making it that much more desirable to use in this low-resource language context. The topic classification of psychological-emotional impact, teaching problems, technology conditions, and

learning problems was successfully achieved with high precision, providing potential practical implications of these models in education policymaking and planning. Furthermore, NLP, with its application, facilitated a smoother, automated process in traditional, time-consuming thematic analysis, providing a replicable, scale-up process for coding large amounts of qualitative data. These findings not only confirm employed models but also vindicate the necessity for the development and employment of language-specific resources for more inclusive, equitable NLP applications. More importantly, this study provides a basis for future research in employing AI-based methods in the analysis of education feedback and stakeholders' narratives, more seriously in language contexts that, until now, have been under-represented in the literature. With education systems evolving in pandemics, evidence-based responses based on structured language models can potentially provide insightful recommendations in developing responsive, evidence-based interventions.

6. 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 modelling 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. ACMTrans. 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. 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.

7. 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.
8. Shibu, P., & Surendran, S. (2023). *Topic Classification of Interviews on Emergency Response: A Machine Learning Approach*. In Proceedings of the 12th International Conference on Data Science, Technology and Applications (DATA 2023), pages 315-322. SCITEPRESS – Science and Technology Publications. https://doi.org/10.5220/0012063600003546 .