Evaluating Language Models for Low-Resource NLP: A Comparative Study of RoBERT and Large Multilingual LLMs

EDI-CRISTIAN BERISHA, University of Twente, The Netherlands

This thesis investigates the performance of RoBERT, a Romanian-language adaptation of the BERT model, in comparison with Gemini, a large language model (LLM) developed by Google, on several Romanian natural language processing (NLP) tasks. While LLMs have demonstrated impressive capabilities across many languages and tasks, their effectiveness in low-resource languages like Romanian remains underexplored. This study addresses this gap by evaluating both RoBERT and Gemini on five key Romanian NLP tasks: sentiment analysis, named entity recognition, topic identification, dialect identification, and offensive language detection.

The models are tested using publicly available Romanian datasets, and their performance is compared using the F1-score as the evaluation metric. The results show that RoBERT outperforms Gemini on tasks that require detailed language-specific knowledge, particularly named entity recognition and dialect identification, while Gemini performs competitively on more general tasks such as sentiment analysis. These findings suggest that, despite the broad generalization abilities of large multilingual models, monolingual models like RoBERT continue to offer important advantages in low-resource language settings, especially when linguistic precision is critical.

Additional Key Words and Phrases: RoBERT, Large Language Models, Gemini, Low-Resource Languages, NLP, Romanian NLP

1 INTRODUCTION

In recent years, language models have become essential tools in natural language processing (NLP), delivering strong performance on tasks such as text classification, sentiment analysis, and named entity recognition. One influential example is **BERT** [6], which has inspired the creation of many language-specific models. Among these is **RoBERT**, a version trained specifically on Romanian data. Romanian is often considered a **low-resource language** [9] in NLP due to the limited availability of annotated datasets, linguistic resources, and domain-specific tools, especially when compared to languages like English, Spanish, or Chinese. While RoBERT is pre-trained on a large Romanian corpus and captures many language-specific patterns, applying it effectively to downstream tasks still requires **task-specific fine-tuning data**. In low-resource settings, obtaining such annotated datasets can be difficult, limiting the full potential of otherwise capable models like RoBERT [11].

At the same time, the field has seen the rise of **large multilingual language models** (LLMs) such as GPT-4 [12], and Google's Gemini [16], which are trained on massive corpora covering many languages. These models are capable of performing a wide range of tasks across different languages without being fine-tuned for each one. Their ability to generalize across multiple languages has raised the question of whether dedicated, language-specific models are still necessary, particularly for languages that are underrepresented in global datasets. Although LLMs like Gemini have shown strong performance across many languages, their effectiveness in low-resource settings remains a topic of active research [14], with some researchers questioning whether such models can adequately support digital equality for underrepresented languages. It is not yet clear whether these general-purpose models can match or surpass the accuracy of smaller, fine-tuned monolingual models like RoBERT when it comes to tasks that require a deep linguistic understanding.

This thesis aims to explore this question by directly comparing the performance of RoBERT and Gemini Flash 2.0 (referred to as Gemini throughout the paper) on a set of core Romanian NLP tasks. These tasks include: sentiment analysis, named entity recognition, topic identification, dialect identification and offensive language detection. The comparison is based on publicly available Romanian datasets, and highlights how each model handles language-specific challenges.

2 RESEARCH QUESTION

This study investigates whether large multilingual language models, such as Google's Gemini, can match or exceed the performance of language-specific models like RoBERT on Romanian NLP tasks. The central question is whether general-purpose models trained across many languages are effective substitutes for fine-tuned monolingual models in low-resource settings where language-specific nuances and limited annotated data pose unique challenges.

To address this, the following research question is formulated:

> *Can large multilingual language models, such as Gemini, achieve performance on Romanian NLP tasks that is comparable to or better than that of a language-specific model like RoBERT in low-resource settings?*

The answer to this question is intended to provide insight into the strengths and limitations of both approaches, helping to clarify whether multilingual models can replace monolingual ones for certain NLP applications in Romanian and potentially other low-resource languages.

3 RELATED WORKS

Since the introduction of **BERT (Bidirectional Encoder Representations from Transformers)** [6], transformer-based models have become the foundation of many advances in natural language processing. BERT is first pre-trained on large general corpora to learn contextual representations of language, and then fine-tuned on smaller, task-specific datasets. This approach enables BERT to capture contextual information in both directions, making it highly effective for tasks such as classification and named entity recognition.

TScIT 43, July 4, 2025, Enschede, The Netherlands

Following the success of BERT, several monolingual variants were developed to improve performance in specific languages. For example, **CamemBERT** [10] was trained for French, **BETO** [4] for Spanish, **AraBERT** [1] for Arabic, **BERTje** [5] for Dutch, and **FinBERT** [17] for Finnish. These models are trained on language-specific corpora and often outperform multilingual models on tasks in their respective languages. In this work, we focus on **RoBERT**, a Romanian BERT model [11]. It was trained entirely on Romanian data using the same transformer-based architecture and demonstrated strong performance on Romanian-specific tasks. The motivation behind RoBERT and similar monolingual models is that language-specific training allows the model to better capture linguistic patterns, structures, and morphology unique to that language which is particularly valuable for languages with limited digital resources, where multilingual models may lack sufficient representation.

In parallel, the emergence of **large-scale multilingual language models (LLMs)** has reshaped the NLP landscape. These models, trained on a mixture of data from many languages, are designed to perform a wide range of tasks without needing fine-tuning for each individual language. Gemini [16], developed by Google DeepMind, is one of the most recent examples of this category. The Gemini model family includes several variants, which are optimized for different performance and efficiency trade-offs. **Gemini Flash 2.0**, used in this study, is a lightweight version designed for speed and low-latency inference. Like other models in the Gemini family, it can handle tasks in a **zero-shot setting**, meaning it can follow task instructions in natural language without prior task-specific training. According to the Gemini Technical Report [16], the model family was trained on a multilingual and multimodal dataset covering over 100 languages, including Romanian. However, the report does not provide evaluations on Romanian NLP benchmarks, and empirical studies on its performance in low-resource language settings remain limited. Moreover, the performance of lightweight, and thus cheaper, multilingual models in truly low-resource environments is often assumed but rarely quantified, especially when it comes to fine-grained linguistic features like dialectal variation. These areas pose specific challenges that may expose limitations in instruction-tuned models not optimized for any particular language.

Meanwhile, Romanian NLP has benefited in recent years from the release of several labeled datasets for core tasks, some notable examples include **RONEC** (**RO**manian **N**amed **E**ntity **C**orpus) [7] for named entity recognition, and **LaRoSeDa** (**La**rge **Ro**manian **Se**ntiment **Da**taset) [15], a sentiment analysis corpus. For offensive language detection, this study uses both the **RO-Offense dataset** [13] and the **ro-fb-offense dataset** [2], which contains annotated comments from Romanian Facebook pages. Additionally, the **MOROCO** (**MO**ldavian and **RO**manian Dialectal **CO**rpus) [3] dataset has been used for dialect identification (distinguishing between Moldavian and Romanian variants) and topic classification. These datasets have enabled the fine-tuning and evaluation of Romanian-specific models like RoBERT across a wide range of NLP tasks. Comprehensive comparisons between specialized and general-purpose models remain scarce in Romanian NLP. The following section outlines the approach taken in this study to address that gap.

4 METHODOLOGY

4.1 Overview

This section describes the methodology used to evaluate and compare the performance of RoBERT and Gemini, on a range of Romanian NLP tasks. Each model is evaluated using publicly available Romanian datasets, and performance is measured using F1-score, a metric that balances precision and recall. RoBERT is fine-tuned individually for each task, while Gemini is evaluated in a zero-shot setting. This section provides details on the models, the datasets, the preprocessing steps, the prompting strategy used for Gemini, and the technical setup used to run the experiments.

4.2 Models

In this study, we use the **RoBERT-base** variant of **RoBERT**, a Romanian-specific model based on the BERT architecture, with approximately 114 million trainable parameters. According to Masala et al. [11] **RoBERT** follows the BERT-base configuration and uses a vocabulary of 38,000 tokens, consistent across all RoBERT variants. We selected this version due to its manageable size and efficiency, which make it well-suited for fine-tuning within a reasonable time frame. **RoBERT** was pre-trained on a Romanian-only corpus containing approximately 2.07 billion words collected from sources such as OSCAR, Wikipedia, and news websites.

The **Gemini** model that was used in this study has a knowledge cut-off in August 2024. It was trained on a large multilingual and multimodal corpus covering over 100 languages, including Romanian, although language-specific performance metrics were not disclosed. **Gemini** was selected for this study due to its accessibility via API, fast response time, and ability to generalize across tasks and languages without retraining.

4.3 Tasks and Datasets

This study evaluates RoBERT and Gemini on five Romanian NLP tasks. These tasks were selected primarily based on the availability of labeled Romanian datasets suitable for benchmarking and to ensure comparability with the original RoBERT paper [11], which focused on similar task categories. Below is a brief overview of each task and the dataset used, the label distribution for each dataset can be found in Appendix A.

Sentiment Analysis

The **LaRoSeDa** dataset contains 15,000 Romanian product and service reviews annotated with a star-based rating system. Only reviews with 1, 2, 4, or 5 stars are included; neutral 3-star reviews were excluded by the dataset creators. Following the dataset's official labeling scheme, reviews rated with 1–2 stars are treated as negative, while those rated with 4–5 stars are treated as positive.

Named Entity Recognition

RONEC is a manually annotated dataset containing 11000 Romanian texts labeled with various entity types, such as persons, organizations, and locations.

Topic Identification
The **MOROCO** dataset includes 27643 text samples from online news sources categorized into six topics: politics, finance, culture, sports, science and tech. This task involves predicting the topic label of each text.

Dialect Identification
The same **MOROCO** corpus is used to distinguish between Moldavian and Romanian dialects. This is a binary classification task, with labels assigned based on the regional origin of the text.

Offensive Language Detection
Two datasets were used for this task, both using the same type of labels. **RO-Offense** includes 12445 user comments from Romanian sports forums, while **Ro-fb-offense** contains 4455 Romanian Facebook comments. The comments were labeled into multiple categories: Other, Profanity, Insult, and Abuse. After fine-tuning, each model was evaluated not only on its own test split but also on the test split of the other dataset to assess cross-dataset generalization. Additionally, a focused evaluation was conducted using only the Other and Profanity classes, as these two categories had relatively similar definitions across the datasets. In contrast, Insult and Abuse differed significantly in how they were defined and annotated, making direct comparison less reliable.

4.4 Data Preprocessing and Implementation

Each dataset was prepared in a way that works with both RoBERT and Gemini, but without making unnecessary changes to the original Romanian text. Because the two models use text differently, RoBERT requires tokenized and labeled inputs for training, while Gemini takes plain text with natural-language instructions, the preprocessing steps were customized for each model.

For **RoBERT**, all datasets were prepared using Hugging Face's Dataset objects to ensure efficient batching and compatibility with the training loop. Input sequence lengths were capped at the 95th percentile of tokenized text lengths for each dataset. This strategy helps reduce memory usage and training time while retaining the vast majority of the content. It also ensures that inputs stay well within the model's maximum input length limit of 512 tokens, beyond which RoBERT cannot process sequences. Labels were mapped to integer class indices as required for classification tasks. To maintain consistency with prior work and ensure comparability, we used the same batch size and learning rate as reported in the original RoBERT paper [11].

Gemini was evaluated in a zero-shot setup using prompts written in natural language. Inputs were simply formatted into task-specific instructions (see Section 4.6), and the model was instructed to return the appropriate label as a number (e.g., 0 or 1). This made it possible to use the outputs directly for evaluation without any additional mapping.

For all tasks, the Romanian diacritics were preserved and the datasets were used in their original train/test splits, as provided by their creators.

4.5 Evaluation strategy

To evaluate model performance across all tasks, this study primarily used the **weighted F1-score**. The F1-score balances precision (the proportion of correct positive predictions) and recall (the proportion of actual positives that were identified), making it a reliable indicator of classification performance.

The **weighted F1-score** computes the F1-score separately for each class and then averages them, giving more weight to classes that occur more frequently in the dataset. This ensures that common classes influence the final score more heavily than rare ones, without completely ignoring minority classes. This metric was used for all binary and multi-class classification tasks, including sentiment analysis, topic identification, dialect identification, and offensive language detection.

For the named entity recognition (NER) task, however, we used the **entity-level F1-score** computed using the *seqeval* library. This evaluation measures the accuracy of predicted entity spans and types, which is standard in sequence labeling tasks like NER.

Each model was evaluated on the designated test split of each dataset:

- **RoBERT** was fine-tuned on the training data for each task and then evaluated on the test set using this metric.
- **Gemini** was evaluated in a zero-shot setting using standardized natural-language prompts (see Section 4.6). The model was instructed to return a class index, so the outputs could be directly compared to ground truth labels without additional processing.

4.6 Prompting Gemini Flash

Since Gemini was evaluated only in a zero-shot setting, task instructions were provided using natural language prompts. The goal was to frame each input as a clear, self-contained instruction so the model could understand what task to perform without requiring fine-tuning. No in-context examples or few-shot prompts were used. The prompts were identical across all examples for a given task, ensuring consistency. Prompt phrasing was based on Google's official Gemini prompt design guidelines [8], which recommend using clear instructions, consistent formatting, and avoiding ambiguity or unnecessary complexity. This approach ensured that the prompts aligned with the model's intended usage and maximized the chance of accurate zero-shot responses.

Sentiment Analysis "You are a sentiment analysis model. Read the Romanian sentence below and classify its sentiment as:
0 = Negative (if the review is critical or low-rated)
1 = Positive (if the review is supportive or high-rated)
Sentence: [text]
Sentiment (0 or 1):"

Named Entity Recognition "Identify the named entity label for each word in this Romanian sentence.

Use labels like O, B-PERSON, I-PERSON, B-ORG, I-ORG, B-GPE, I-GPE, B-LOC, I-LOC, B-NAT_REL_POL, I-NAT_REL_POL, B-EVENT,

I-EVENT, B-LANGUAGE, I-LANGUAGE, B-WORK_OF_ART, I-WORK_OF_ART, B-DATETIME, I-DATETIME, B-PERIOD, I-PERIOD, B-MONEY, I-MONEY, B-QUANTITY, I-QUANTITY, B-NUMERIC, I-NUMERIC, B-ORDINAL, I-ORDINAL, B-FACILITY, I-FACILITY.
Sentence: [text]
Format your answer as:
word1: label1
word2: label2
...
Only include words from the sentence. "

Topic Identification "What theme does the following romanian article talk about? choose one of the following categories:
0 - culture
1 - finance
2 - politics
3 - science
4 - sports
5 - tech
Text: [text]
Respond only with the number corresponding to one of the above categories.
No other comment. "

Dialect Identification "You are a dialect identification model. Read the Romanian sentence below and classify its dialect as:
0 = Moldavian (if the text is written in Moldavian dialect)
1 = Romanian (if the text is written in Romanian dialect)
Sentence: [text]
Dialect (0 or 1):"

Offensive Language Detection 1 (Ro-Offense Dataset) "Classify the following Romanian text into one of the categories below and respond only with the corresponding number:
0 - Other: Neutral or non-offensive content.
1 - Profanity: Comments containing curse words or offensive words that are not directed at a person or a group and do not disparage certain minority groups. These messages are not intended to hurt anyone but contain profane words, and most are impersonal expressions of grievance.
2 - Insult: Comments meant to offend certain individuals or a group while ascribing negative qualities or deficiencies. These messages convey the feeling of contempt or disrespect towards their target. We included here all allusions to reduced intellectual capacity, barring any reference to mental health or disability. Additionally, this category included most sexual insults that do not imply violence or forced sexual acts.
3 - Abuse: Any type of threat, violence, death, or wishes of sickness. This language ascribes an undesirable social identity that is either judged negatively by society or perceived in a negative light by the majority. Dehumanizing and disparaging language is also classified as ABUSE. Identifying the target as a member of a sexual minority, having disabilities, or labeling him/her as suffering from various mental health issues harms not only the target but also the mentioned minority groups, by feeding into the stigma surrounding these groups. Other stigmatized groups such as sex workers, drug abusers, or homeless people also fall into this category. Any racist, xenophobe, or chauvinist comment.
Text: [text]
Answer with only one number: 0, 1, 2, or 3. "

Offensive Language Detection 2 (ro-fb-offense) "Classify the following Romanian text into one of the categories below and respond only with the corresponding number:
0 - Other: Neutral or non-offensive content.
1 - Profanity: Comment that is not targeted at an individual or a group, but contains swear words or profane expressions.
2 - Insult: Comparison to an animal, insulting expressions without swear words, anger and contempt towards the target, other insults.
3 - Abuse: Racist comments, assigning a group negatively perceived in society, sexist comments / sexual harassment, wishing someone deadly diseases, death wishes, cursing the targe.
Text: "text"
Answer with only one number: 0, 1, 2, or 3. "

For the **offensive language detection** task, the definitions of each label used in the Gemini prompts were extracted directly from the dataset documentation provided by the creators of the original corpus. This approach ensured that the prompts closely aligned with the original annotation guidelines, maintaining consistency between how the data was labeled and how the model was instructed to classify each instance.

4.7 Environment

All experiments involving RoBERT were carried out using **JupyterLab**, hosted on computing infrastructure provided by the **University of Twente**. Model fine-tuning and evaluation were implemented using the *Hugging Face Transformers* and *Datasets* libraries. Training was executed using the *transformers.Trainer* API, and evaluation was performed with the *evaluate* library, using the F1-score as the primary metric. Training times ranged from several minutes to several hours per task, depending on dataset size.

Prompts for Gemini were submitted using the Google Gemini API. The API was accessed using Python scripts, and responses were parsed automatically to extract the class predictions, which were then compared to ground truth labels.

5 RESULTS

Table 1. F1-scores for each NLP task across the two models

Task	RoBERT	Gemini
Sentiment Analysis	0.9553	**0.9590**
Named Entity Recognition	**0.8720**	0.4132
Topic Identification	**0.8631**	0.7758
Dialect Identification	**0.9633**	0.4132
Offensive Language Detection 1	**0.8170**	0.6802
Offensive Language Detection 2	**0.8615**	0.7485

Table 1 presents the **F1-scores** for each Romanian NLP task across the two evaluated models. Overall, RoBERT outperforms Gemini

on tasks requiring a good understanding of the nuanced details of the Romanian language. However, Gemini shows strong results on certain general classification tasks, despite being used in a zero-shot setting. This section provides a detailed breakdown of each task and discusses the implications of the observed performance trends.

5.1 Sentiment Analysis

Both models performed strongly on the sentiment analysis task, with **Gemini** and **RoBERT** achieving nearly identical scores (0.9590 and 0.9553, respectively). This task involved deciding whether a review was positive or negative. Since the reviews were usually clear in expressing opinions, it was easier for both models to understand the meaning. This shows that even without being fine-tuned, Gemini can handle basic tasks like this quite well in Romanian.

5.2 Named Entity Recognition

RoBERT clearly outperformed Gemini on this task, scoring 0.8720 compared to Gemini's 0.4132. Named entity recognition (NER) requires identifying specific types of information, such as names of people, places, or organizations, within a sentence, which often depends on understanding subtle linguistic cues and sentence structure. Unlike sentiment analysis, this task is more sensitive to word boundaries and contextual clues. Gemini, evaluated in a zero-shot setting, struggled with this level of precision. RoBERT, on the other hand, was fine-tuned specifically for this task using Romanian data, which enabled it to better capture the patterns necessary for accurate entity recognition.

5.3 Topic Identification

RoBERT also performed better than Gemini on this task (0.8631 vs. 0.7758), although the difference was smaller than in NER. In this case, the model had to figure out what the provided piece of text was about (e.g., politics, sports, etc.). Gemini achieved strong results on this task, likely because the task focuses more on general understanding rather than exact word usage. Still, RoBERT had an advantage by being trained specifically on Romanian text and topic categories.

5.4 Dialect Identification

This was the task with the biggest difference between the two models. RoBERT scored 0.9633, while Gemini only got 0.4132. The goal here was to tell whether a sentence was written in Moldavian or standard Romanian. These differences are very subtle and often depend on regional words or spelling variations. Gemini did not perform well on this task, likely because, even though it may have been exposed to both Romanian and Moldavian texts during training, it was not specifically trained to distinguish between them. RoBERT, however, was fine-tuned on a dataset built for this exact purpose, which helped it perform extremely well.

We examined the following Moldavian sentence that Gemini mislabeled, while RoBERT correctly identified the dialect: *"Ufologii au relatat despre legătura dintre furtună și observarea obiectelor zburătoare neidentificate, transmite NE* [1] *. Potrivit experților, navele de zbor ale reprezentanților civilizațiilor extraterestre dispun de protectoare, ceea ce le permite să nu fie văzute . Cu toate acestea, uneori, o astfel de protecție poate eșua . De exemplu, acest lucru se poate întîmpla în timpul unei furtuni, cînd protecția navei NE scade . Un inginer a scris o carte despre o pușcărie secretă pentru extratereștriNE dată dovada acestei versiuni a fost observată în NE în timpul uraganului „NE . Ufologii fac referire la fotografiile NE care au surprins obiecte neidentificate lîngă vîrtej . Entuziaștii s - au grăbit să îi numească nave ale extraterestre . În același timp, ei explică calitatea joasă a fotografiilor prin autoprotecția extratereștrilor, realizată cu ajutorul undelor electromagnetice ."*

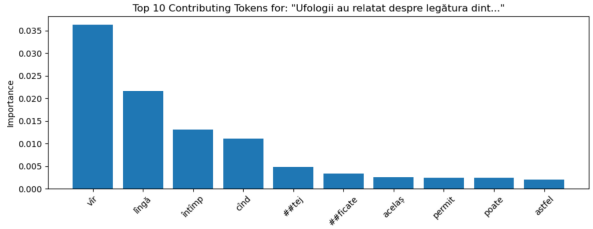

Fig. 1. Token importance for dialect prediction on a Moldavian sentence.

Figure 1 illustrates the most important tokens that contributed to RoBERT's correct prediction of the dialect in the example sentence. We used the **Integrated Gradients** method from the *Captum* library to compute token-level attribution scores. This technique measures the contribution of each input token to the model's prediction by integrating gradients along a path from a baseline input to the actual input embeddings. Only the top 10 tokens (by absolute attribution score) are shown in the figure for clarity. Tokens such as *vîr*, *lîngă*, and *întîmp* carry strong regional signals characteristic of the Moldavian dialect. These tokens appear in words like "vîrtej," "lîngă," and "întîmplător," which use the letter *î* in the middle of the word, a spelling convention more typical in Moldavian usage, as opposed to the standard Romanian convention of using *â* in that position. This suggests that RoBERT has learned to rely on subtle orthographic cues to distinguish between dialects.

5.5 Offensive Language Detection

This task used two datasets: Ro-Offense and ro-fb-offense. RoBERT scored 0.8170 and 0.8615 on them, while Gemini scored 0.6802 and 0.7485. Although Gemini's scores were decent, RoBERT was more accurate in both cases. This task involves detecting when a text contains harmful or offensive language. Understanding this in Romanian requires knowing how people actually speak online, including slang or indirect insults. RoBERT learned these patterns during training, which gave it an advantage.

To further investigate generalization, we tested how each RoBERT model performed on the other dataset. The Ro-Offense model scored 0.6804 on ro-fb-offense, while the Ro-fb-offense model scored 0.5983 on Ro-Offense. However, direct comparisons are complicated by differences in how the two datasets define categories like Insult and Abuse. To address this, we also conducted a focused evaluation using only the Other and Profanity labels, which had more consistent definitions. In this setting, the Ro-Offense model scored 0.8618,

[1] NE refers to named entities. Certain names and locations were masked in the original dataset to prevent the model from learning dialect associations from specific entities.

and the Ro-fb-offense model scored 0.8329, suggesting stronger agreement when label interpretation was better aligned.

5.6 Summary

To summarize, RoBERT performed better overall, especially on tasks where it helped to know Romanian grammar, spelling, or regional expressions. Gemini was most successful on general tasks like sentiment analysis and topic identification, but it struggled more when the task required deeper language knowledge. This suggests that while large models like Gemini can be useful, language-specific models like RoBERT still offer important advantages in low-resource languages like Romanian.

6 DISCUSSION

The results of this study show that RoBERT performs **better** than Gemini on most Romanian NLP tasks, especially when the task requires detailed understanding of the Romanian language. Gemini performs well on general tasks, such as sentiment analysis and topic identification, but struggles with more specific or subtle language features.

The most significant performance gaps appeared in dialect identification and named entity recognition. These are tasks where it helps to know the grammar, spelling patterns, and regional differences of Romanian. For example, in the dialect task, RoBERT was able to recognize patterns like the use of "î" instead of "â" in Moldavian texts. This kind of detail is not easy for general language models like Gemini to detect, especially when they are used in a zero-shot setting and have not been fine-tuned for Romanian. A similar pattern emerged in named entity recognition, where RoBERT's familiarity with Romanian syntax and token boundaries gave it a substantial advantage.

Gemini performed relatively well on sentiment analysis and topic identification, even slightly outperforming RoBERT in sentiment analysis. These tasks depend more on understanding the overall meaning or tone of the text than on specific grammatical or lexical details. Because Gemini was trained on a wide variety of multilingual data, it is able to generalize well to tasks that rely on broad semantic patterns.

Offensive language detection yielded nuanced results. RoBERT consistently outperformed Gemini in terms of weighted F1-score on both datasets, demonstrating strong task-specific performance when fine-tuned. Cross-dataset evaluation showed that RoBERT's performance dropped when applied to the other dataset, but this may be due to differences in how each dataset defines categories such as Insult and Abuse, as well as differences in domain and language use. To reduce the impact of label definition mismatch, a focused evaluation was performed using only the Other and Profanity categories, which showed more consistent performance.

Gemini was evaluated only in a zero-shot setting. While we cannot determine whether it was exposed to similar data during training, its lower scores suggest that it may not handle the nuanced and context-dependent nature of offensive language in Romanian as well as a fine-tuned model.

Overall, the results highlight how model performance in offensive language detection is influenced not only by architecture and training setup, but also by the consistency and clarity of annotation guidelines across datasets. The task of offensive language detection proved to be one of the most difficult in the study. Offensive language is highly contextual, culturally embedded, and often ambiguous. Even as a native Romanian speaker, I found several examples difficult to label or agree with. Many sentences blurred the lines between categories, such as combining insults with abusive or profane elements. This kind of overlap makes it difficult to assign a single, correct label and shows that the challenge lies not just in modeling, but in the annotation process itself. Disagreements between annotators, as well as subjective interpretations of terms like "abuse" or "insult," introduce uncertainty into the datasets. Informal and nonstandard language, including slang, abbreviations, and sarcasm, also acts as a confounding variable, making both annotation and classification more difficult. This ambiguity affects evaluation fairness and model learning. Improving the clarity of annotation guidelines, allowing overlapping or multi-label annotations, and defining more specific categories (e.g., racism, sexism, threats) could help build better models for offensive language detection in Romanian and potentially other low-resource languages.

6.1 Limitations

While this study provides useful insights into the performance of RoBERT and Gemini on Romanian NLP tasks, there are several limitations to consider. First, Gemini was only evaluated in a zero-shot setting using single-instruction prompts. It is possible that few-shot prompting or task-specific fine-tuning could significantly improve its performance, particularly for complex classification tasks.

Second, although RoBERT performed well overall, it was pretrained and fine-tuned on datasets that vary in domain. For tasks like sentiment analysis, named entity recognition, topic identification, and dialect identification, the data comes mostly from formal sources such as product reviews or news articles. This may limit generalization to informal or spoken Romanian. However, in the case of offensive language detection, both Ro-Offense and ro-fb-offense contain informal, user-generated content from forums and social media, making them more representative of real-world, unstructured text. Even so, RoBERT's performance on these datasets may still be affected by differences in writing style, slang, or domain-specific phrasing.

Another limitation concerns the datasets used for offensive language detection. While the other tasks in this study involved clearly defined and objective labels, offensive language detection is inherently more subjective. Terms such as insult, profanity, and abuse are difficult to define with precision, and the boundaries between them are often unclear. Offensive content can express multiple types of harm simultaneously, and many examples naturally overlap across categories. For instance, a sentence might include both a personal insult and a dehumanizing comment, making it challenging to assign a single, correct label. This becomes especially problematic when different types of offensive language are meant to be handled differently, such as when moderation systems or legal frameworks apply distinct responses to insults versus hate speech.

This ambiguity is not only difficult for models to handle, it also poses a problem for human annotators. In practice, different annotators may interpret the same sentence differently based on personal, cultural, or contextual factors. Without well-documented annotation guidelines or reported inter-annotator agreement scores, it is difficult to assess the consistency of labeling across the dataset. These issues introduce uncertainty into both training and evaluation, and model performance on this task should therefore be interpreted with caution.

Furthermore, this study only tested Gemini Flash 2.0, which is one of the smaller and faster variants in the Gemini model family. Due to API access limitations and cost restrictions, more powerful models such as Gemini Ultra or GPT-4 were not available for evaluation. In addition, the training data used for large proprietary models like Gemini is not publicly disclosed, making it difficult to assess whether the model was exposed to similar datasets during pre-training. This lack of transparency complicates direct comparisons and raises questions about potential data leakage or unseen advantages. As a result, the comparison does not fully represent the current upper bounds of large language model performance.

6.2 Future work

This study highlights several directions for future research. First, further exploration is needed into the capabilities of more advanced multilingual models, future work could include larger and more powerful models, such as Gemini Ultra or GPT-4, to assess whether the performance gap with RoBERT remains when stronger LLMs are used.

Second, the impact of few-shot prompting or in-context learning for LLMs could be investigated. This study focused on a zero-shot setting for consistency and simplicity, but it is possible that providing examples or task-specific context could improve results, especially for more complex tasks like NER or offensive language detection.

Third, while RoBERT performed well on formal written text, future research could examine its performance on informal, conversational, or social media data. Fine-tuning RoBERT on more informal text could help it work better in real-world situations, especially since casual language is common in online conversations in low-resource languages.

Finally, offensive language detection remains a complex task, even when datasets include clear label definitions. Many offensive messages naturally belong to more than one category. Even with well-defined criteria, human annotators may still disagree, showing the subjective nature of the task. In the future, datasets that allow multiple labels per message could help models better capture these overlaps. It could also be useful to tag specific types of offensive language, such as racism, sexism, or threats, to make detection more precise and socially relevant. Exploring multi-label classification techniques and improving evaluation methods to handle ambiguity would be valuable next steps in advancing this area.

6.3 Answering the Research Question

This study set out to explore the following research question:

Can large multilingual language models, such as Gemini, achieve performance on Romanian NLP tasks that is comparable to or better than that of a language-specific model like RoBERT in low-resource settings?

The experimental results show that Gemini performs reasonably well on general tasks like sentiment analysis and topic identification. These tasks focus more on understanding overall meaning and less on language-specific structure, making them more suitable for a zero-shot multilingual model.

However, Gemini struggled significantly with tasks that require deeper linguistic understanding, such as dialect identification, named entity recognition, and offensive language detection. In contrast, RoBERT consistently performed better across all tasks, particularly those that rely on grammatical cues, spelling conventions, or region-specific language.

Therefore, in response to the main research question: multilingual language models like Gemini can perform competitively on some Romanian NLP tasks, but they do not yet match the performance of a language-specific model like RoBERT across the board. RoBERT remains the more reliable option, especially for tasks requiring fine-grained, culturally and linguistically informed understanding.

6.4 Practical Implications

Beyond performance, the two models compared in this study differ significantly in how they can be used in real-world applications. RoBERT consistently outperformed Gemini across most tasks, demonstrating its effectiveness when fine-tuned for Romanian-specific data. In addition to strong performance, RoBERT can be downloaded and run locally on personal hardware, offering full control over the model and data. This makes it more suitable for processing large volumes of text efficiently, especially in cases where privacy, repeatability, or offline usage is important. For organizations working with sensitive data or deploying NLP at scale in Romanian, RoBERT offers a practical and cost-effective solution.

Gemini Flash, on the other hand, is easy to use and accessible through an API, requiring no local hardware or setup. This makes it suitable for quick prototyping or integration into cloud-based workflows. However, it depends on internet access, may have usage limits, and is subject to the availability and pricing of commercial APIs.

7 CONCLUSION

This study compared the performance of RoBERT, a Romanian-specific transformer model, with Gemini , a multilingual large language model, across five core Romanian NLP tasks: sentiment analysis, named entity recognition, topic identification, dialect identification, and offensive language detection. The goal was to evaluate whether a general-purpose, zero-shot multilingual model can match the performance of a fine-tuned monolingual model in a low-resource language setting.

The results show that RoBERT outperformed Gemini in most tasks, especially those requiring detailed linguistic understanding, such as dialect identification and named entity recognition. Gemini performed reasonably well on tasks that rely more on broad semantic understanding, such as sentiment and topic classification, but

struggled with tasks that involved regional variation, syntax, or nuanced definitions of offensive language.

These findings suggest that while multilingual LLMs have made significant progress and offer ease of use through prompt-based evaluation, they do not yet fully replace monolingual models when linguistic precision is important. RoBERT remains a reliable and efficient choice for Romanian NLP, particularly in offline or large-scale processing scenarios.

At the same time, the performance of Gemini Flash in zero-shot conditions demonstrates that LLMs can be viable tools for certain Romanian NLP tasks, especially where resources for fine-tuning are limited.

REFERENCES

[1] Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based Model for Arabic Language Understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020.*

[2] Gabriel-Razvan Busuioc, Andrei Paraschiv, and Mihai Dascalu. 2022. FB-RO-Offense – A Romanian Dataset and Baseline Models for detecting Offensive Language in Facebook Comments. In *International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) 2022.*

[3] Andrei M. Butnaru and Radu Tudor Ionescu. 2019. MOROCO: The Moldavian and Romanian Dialectal Corpus. arXiv:1901.06543 [cs.CL] https://arxiv.org/abs/1901.06543

[4] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020.*

[5] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. arXiv:1912.09582. http://arxiv.org/abs/1912.09582

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[7] Stefan Daniel Dumitrescu and Andrei-Marius Avram. 2019. Introducing RONEC–the Romanian Named Entity Corpus. *arXiv preprint arXiv:1909.01247* (2019).

[8] Google. 2024. Prompt design strategies — Gemini API. https://ai.google.dev/gemini-api/docs/prompting-strategies

[9] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[10] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.645

[11] Mihai Masala, Stefan Ruseti, and Mihai Dascalu. 2020. RoBERT – A Romanian BERT Model. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 6626–6637. https://doi.org/10.18653/v1/2020.coling-main.581

[12] OpenAI, Josh Achiam, Steven Adler, and Sandhini Agarwal et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[13] ReaderBench Team. 2022. Ro-Offense: RO-Offense: A Novel Romanian Dataset for Offensive Language in Online Comments. https://github.com/readerbench/ro-offense. Accessed: 2025-05-01.

[14] Georg Rehm, Annika Grützner-Zahn, and Fabio Barth. 2025. Are Multilingual Language Models an Off-ramp for Under-resourced Languages? Will we arrive at Digital Language Equality in Europe in 2030? arXiv:2502.12886 [cs.CL] https://arxiv.org/abs/2502.12886

[15] Anca Maria Tache, Mihaela Gaman, and Radu Tudor Ionescu. 2021. Clustering Word Embeddings with Self-Organizing Maps. Application on LaRoSeDa – A Large Romanian Sentiment Data Set. *ArXiv* (2021).

[16] Gemini Team, Rohan Anil, Sebastian Borgeaud, and Jean-Baptiste Alayrac et al. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[17] Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: BERT for Finnish. arXiv:1912.07076 [cs.CL] https://arxiv.org/abs/1912.07076

A DATASET LABEL DISTRIBUTIONS

Table 2. Label distribution in the LaRoSeDa dataset

Label	Train	Test	Total
Postive	6000	1500	7500
Negative	6000	1500	7500
Total	12000	3000	15000

Table 3. Label distribution in the RONEC dataset

Label	Train	Test	Total
PERSON	19167	4230	23397
GPE	8193	1728	9921
LOC	1824	373	2197
ORG	5688	1312	7000
LANGUAGE	342	73	415
NAT_REL_POL	3673	781	4454
DATETIME	6960	1625	8585
PERIOD	862	197	1059
QUANTITY	1161	246	1407
MONEY	1041	224	1265
NUMERIC	5734	1187	6921
ORDINAL	1377	304	1681
FACILITY	840	173	1013
WORK_OF_ART	1157	263	1420
EVENT	826	169	995
Total	58845	12885	71730

Table 4. Topic label distribution in the MOROCO dataset

Label	Train	Test	Total
Culture	1484	404	1888
Finance	5522	1506	7028
Politics	5910	1612	7522
Science	1890	515	2405
Sports	3899	1064	4963
Tech	3014	823	3837
Total	21719	5924	27643

Table 5. Dialect label distribution in the MOROCO dataset

Label	Train	Test	Total
Moldavian	9968	2719	12687
Romanian	11751	3205	14956
Total	21719	5924	27643

Table 6. Label distribution in the Ro-Offense dataset

Label	Train	Test	Total
Other	3649	898	4547
Profanity	1294	331	1625
Abuse	2768	684	3452
Insult	2242	579	2821
Total	9953	2492	12445

Table 7. Label distribution in the ro-fb-offense dataset

Label	Train	Test	Total
Other	2121	552	2673
Profanity	147	32	179
Abuse	668	166	834
Insult	628	141	769
Total	3564	891	4455