Detecting Suicidal Ideation on Social Media Using Large Language Models with Zero-Shot Prompting

Golnaz Nikmehr[1][a], Aritz Bilbao-Jayo[1][b], Aron Henriksson[2][c] and Aitor Almeida[1][d]
[1]*Deustotech - University of Deusto, Bilbao, Spain*
[2]*Stockholm University, Stockholm, Sweden*
{*golnaz.nikmehr, aritzbilbao, aitor.almeida*}*@deusto.es, aronhen@dsv.su.se*

Keywords: Suicidal Ideation Detection, Natural Language Processing, Large Language Models, Prompting.

Abstract: Detecting suicidal ideation in social media posts using Natural Language Processing (NLP) and Machine Learning has become an essential approach for early intervention and providing support to at-risk individuals. The role of data is critical in this process, as the accuracy of NLP models largely depends on the quality and quantity of labeled data available for training. Traditional methods, such as keyword-based approaches and models reliant on manually annotated datasets, face limitations due to the complex and time-consuming nature of data labeling. This shortage of high-quality labeled data creates a significant bottleneck, limiting model fine-tuning. With the recent emergence of Large Language Models (LLMs) in various NLP applications, we utilize their strengths to classify posts expressing suicidal ideation. Specifically, we apply zero-shot prompting with LLMs, enabling effective classification even in data-scarce environments without needing extensive fine-tuning, thus reducing the dependence on large annotated datasets. Our findings suggest that zero-shot LLMs can match or exceed the performance of traditional approaches like fine-tuned RoBERTa in identifying suicidal ideation. Although no single LLM outperforms consistently across all tasks, their adaptability and effectiveness underscore their potential to detect suicidal thoughts without requiring manually labeled data.

1 INTRODUCTION

The World Health Organization (WHO) reports that approximately 800,000 people die by suicide each year, making it one of the leading causes of death worldwide (World Health Organization, 2021). Among people aged 5-29, three of the top five causes of death are injury-related: road traffic accidents, homicide, and suicide (World Health Organization, 2022). Suicide rates differ widely across countries and regions, influenced by factors such as mental health, age, gender, and geographic location. Both the WHO and the American Association of Suicidology (AAS) (American Association of Suicidology, 2023) define suicidal ideation as "thinking about, considering, or planning suicide," encompassing both passive thoughts of death without specific plans and active thoughts involving plans or intent. The AAS emphasizes that suicidal ideation is complex, often arising from a combination of factors like mental health conditions, trauma, and stress.

An increasing number of individuals are using social media platforms like Twitter and online forums such as Reddit to express their emotions and feelings. As a result, analyzing these platforms has become crucial for identifying suicidal ideation. Early detection of such thoughts can enable timely diagnosis and treatment. In this regard, Machine Learning and Natural Language Processing (NLP) are playing a key role in automating the detection of suicidal ideation.

The literature review highlights a major limitation across traditional classification methods and deep learning approaches: the requirement for extensive labeled training data. In their study, Kumar et al. (Kumar et al., 2020) obtained optimal results by using a Random Forest classifier combined with VADER sentiment analysis (Hutto and Gilbert, 2014) and word embeddings to detect suicidal ideation on Twitter. Hutto and Gilbert calculate sentiment scores by summing the ratings of lexicon words, categorizing them as negative, positive, or neutral, with the assumption that sentences labeled as negative or neutral may contain suicidal thoughts.

[a] https://orcid.org/0000-0001-6197-9935
[b] https://orcid.org/0000-0001-7743-6652
[c] https://orcid.org/0000-0001-9731-1048
[d] https://orcid.org/0000-0002-1585-4717

Nikmehr, G., Bilbao-Jayo, A., Henriksson, A. and Almeida, A.
Detecting Suicidal Ideation on Social Media Using Large Language Models with Zero-Shot Prompting.
DOI: 10.5220/0013283400003938

In *Proceedings of the 11th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2025)*, pages 259-267
ISBN: 978-989-758-743-6; ISSN: 2184-4984

In the deep learning aspect, fine-tuned text classification involves pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers) to specific classification tasks by further training them on labeled data from the target domain. BERT, introduced by Devlin et al (Devlin et al., 2018), transformed NLP by pre-training on a large corpus with a masked language modeling objective and then fine-tuning specific tasks, resulting in state-of-the-art performance across various benchmarks. During fine-tuning, the model learns task-specific variations by adjusting its weights based on the labeled examples, leading to superior performance compared to models that rely on pre-training. This approach has been extended in models like RoBERTa (Liu et al., 2019), which optimizes BERT's training methodology for even better results. Fine-tuning enables these models to perform well in text classification tasks across various domains. Specifically in (Haque et al., 2020) present a Transformer model is having pre-trained language models for detecting suicidal ideation.

The main challenge is the time-intensive, labor-heavy task of manually labeling data for each social media platform. Furthermore, as noted in the literature, it remains uncertain how well a model trained on one platform can generalize to others. Our focus is on cases where training data is available from an out-of-domain platform but is lacking in the target (in-domain) platform. Potential strategies include fine-tuning models on out-of-domain data for application in the target domain, using keyword-based methods, or utilizing large language models (LLMs) with zero-shot prompting to automatically label data. This approach reduces reliance on specific datasets or platforms, offering a more scalable and adaptable solution.

LLMs have the capability to handle various applications such as general natural language tasks (Chang et al., 2024). In terms of natural language processing tasks, we focus on using LLMs as a classification tool to detect suicidal ideations in social networks. Besides the classic “pre-train and fine-tune” samples, by way of in-context learning, where one can use a text or template known as a prompt to strongly guide the generation to output answers for desired tasks, thus beginning an era of “pre-train and prompt” (Liu et al., 2023). Zero-shot prompting for text classification tasks uses the capabilities of large pre-trained language models, such as GPT-3 (Floridi and Chiriatti, 2020) and T5 (Raffel et al., 2019), to classify text without explicit task-specific training. In (Brown et al., 2020) Brown et al. demonstrated the effectiveness of GPT-3 in such scenarios, where prompt engineering plays a crucial role in guiding the model to understand and perform the classification task.

We investigate and demonstrate the feasibility of using LLMs to classify social media posts based on their relevance to suicide or the presence of suicidal ideation. Additionally, we apply various methods to accomplish this objective. The primary contributions of this paper are:

- We explore the use of LLMs detection of suicidal ideation in Reddit posts, specifically in a zero-shot prompting setting that does not rely on access to labeled data.
- We experiment with two different types of prompts (question-based and description-based) and model sizes (8B, 70B) and show that LLM prompting outperforms a keyword-based approach and a RoBERTa model fine-tuned on data from another social media.
- We investigate three different classification setups based on different granularity of the target classes. The experiments were conducted using two approaches: direct classification into three categories (’Suicidal’, ’Related’, and ’Unrelated’) and a 2-step classification process.

The paper follows this structure: in Section Related Work, we provide an overview of relevant literature. The Dataset section details the dataset used, including an analysis of its contents. Section Methodology discusses the models and how the experiments were conducted. The Results section presents and analyzes our findings. Finally, the Conclusion Section offers concluding remarks.

2 RELATED WORK

This section introduces the baseline in suicidal ideation detection in social networks such as classical machine learning approaches and recently the use of pre-trained models. Regarding the datasets used in the literature on suicidal ideation detection, including social networks such as Twitter, Reddit, and Facebook and Forums like Sina Weibo, often manually annotated. Some of these papers apply specific lexicons containing phrases or keywords indicative of suicidal ideation. For instance in (Aldhyani and Alshebami, 2022) the identification of suicidal posts was executed through a keyword-based search in Reddit, involving terms such as *suicide*, *kill myself*, and *end my life*. Also, the dataset was annotated according to the subreddit of each post. In Figure 1, we analyze the distribution of platforms utilized for suicidal ideation detection in papers published from 2018 to 2023.

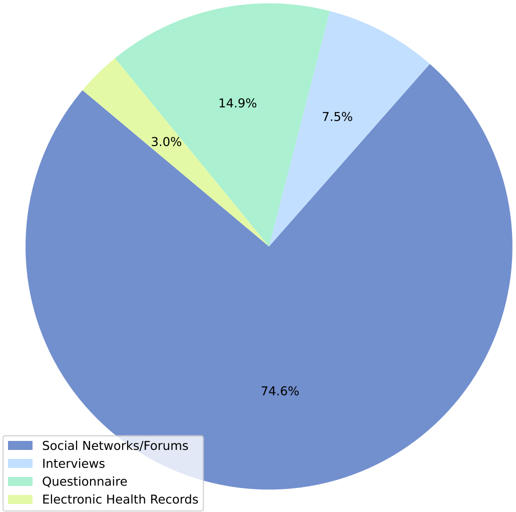


Figure 1: Distribution of platforms employed in papers published from 2018 to 2023.

In basic methods for detecting suicidal ideation, Kumar et al. claim that Random Forest (RF) (Breiman, 2001) as the classical machine learning algorithm, compared to other methods like Logistic Regression, shows higher accuracy (0.996) in identifying tweets indicative of suicidal ideation on Twitter (Kumar et al., 2020). The authors initiate data collection on the Twitter platform by filtering keywords such as 'die,' 'to die,' 'suicide,' 'kill myself,' and 'end my life,' among others. Additionally, they also use n-grams as search keywords to better identify tweets with suicidal thoughts. Tweets with these keywords were then labeled as expressing suicidal ideation.

In a study by (Haque et al., 2020), researchers introduce a Transformer-based method for detecting suicidal ideation in social media posts using pre-trained language models. They use a large dataset of 3,549 texts indicating suicidal thoughts from Reddit, and several non-suicidal texts. The model has three main parts: a data layer, an embedding layer, and a classification layer. In the data layer, they preprocess the suicidal texts by expanding abbreviations and removing URLs. The embedding layer uses pre-trained language models to convert the texts into numerical vectors. The classification layer uses a neural network to classify the texts as suicidal or non-suicidal.

The results show that Transformer-based models outperformed other deep neural network models, with higher accuracy, recall, precision, and F_1-score. The RoBERTa model (Liu et al., 2019), a pre-trained BERT model, was the best across all metrics. The study concludes that the Transformer-based approach with pre-trained language models is an effective tool for detecting suicidal ideation in social media posts.

3 DATASETS

3.1 Data Collection

For our study, we required a dataset of texts containing suicidal ideation. To collect this data, we selected Reddit and a forum called SanctionSuicide as our sources.

3.1.1 Reddit

In this study, Reddit serves as our target data source. To collect Reddit data, we utilized the Reddit API to search across all subreddits using a set of 25 keywords and phrases. Given the vast number of posts on this platform, we needed to limit the data collection to relevant content. Therefore, we decided to use specific keywords to filter the posts. The keywords are chosen from keywords that Ramit et al. (Sawhney et al., 2018) mention in their research. We present these keywords in Table 1. In most previous work in this area, researchers have collected data from specific subreddits like "SuicideWatch" leading to potential bias in the dataset.

Table 1: Keywords and Phrases with Suicidal Intent.

suicidal	suicide
my suicide letter	kill myself
can't go on ready to jump	slit my wrist
cut my wrist	my suicide note
want to die	sleep forever
slash my wrist	wanna die
wanna suicide	commit suicide
take my own life	thoughts of suicide
I wish I were dead	suicide ideation
suicide plan	end my life
never wake up	tired of living
nothing to live for	go to sleep forever
ready to die	

After removing duplicates and filtering for posts that contain specific keywords and were posted in 2023, we have 172 posts. We then extracted all the comments from these posts, limiting the number of comments between 0 and 15, resulting in 298 comments (without considering keywords in comments). These comments do not contain specific keywords, unlike the posts, and are therefore more general in nature. We manually annotated the Reddit data based on our previously defined criteria: if a post mentions "thinking about, considering, or planning suicide," it is labeled as 'Suicidal' Posts related to suicide in general or previous attempts are labeled as 'Related' Posts not fitting these categories are labeled as 'Unrelated'.

3.1.2 SanctionSuicide

SanctionSuicide is a thread-based forum where users create threads and others comment on them. Despite being a valuable source of genuine expressions from users with suicidal intent, it hasn't been discussed in recent research papers. We selected this forum for data collection to train our pre-trained model because, unlike Reddit, it does not impose restrictions on the amount of data we can gather. We collected 77,002 threads using Selenium, which consisted of Titles, URLs, Views, Replies, and Users. Since the dataset was extensive, it was impractical to label all of it manually. To address this, we used BERTopic (Grootendorst, 2022) to categorize the titles into five distinct topics. We then randomly selected threads from one of these topics for annotation. Ultimately, we manually labeled 5,010 posts from 334 threads on this forum into two categories based on the definitions provided: 'Suicidal' and 'Unrelated'. Table 2 presents the distribution of the number of posts and comments in these two datasets.

Table 2: Class label distribution of the dataset.

Dataset	Suicidal	Related	Unrelated
Reddit-Post	82 (48%)	65 (37%)	25 (14%)
Reddit-Comments	6 (2%)	43 (14%)	249 (84%)
SanctionSuicide	897 (18%)	-	4113 (82%)

3.2 Data Analysis

Since Reddit data is our target for the study's experiments, we analyzed the labeled data to gain a deeper understanding. Using keywords and phrases from Table 1, we examined their distribution in Reddit posts for each label. We noticed that even the 'Unrelated' class contains these keywords, which could make distinguishing these posts from 'Suicidal' ones challenging and it's a limitation for keyword-based approaches. The distribution is shown in Figure 2. Additionally, even though we didn't use specific phrases to collect comments, we found that keywords still appear in them.

Not all posts where a person expresses suicidal intent have comments related to suicide, and posts related to suicide may include comments about suicidal thoughts. This creates challenges for detection methods that rely solely on word patterns, potentially leading to inaccurate results. It is crucial to distinguish between these cases, as many methods struggle with this distinction. Also, analysis of comment distribution across post labels reveals that many comments on 'Suicidal' posts are labeled as 'Unrelated,' as shown in Figure 3. This suggests that most comments on 'Suicidal' posts do not directly address suicide, making relevant comment data relatively scarce.

4 METHODS

4.1 Data Pre-Processing

Our content, sourced from social media, includes URLs, mentions, usernames, and special characters. The first step is text cleaning, which involves identifying mentions and usernames (marked by '@') and removing extra spaces and lines. We focus only on English posts.

4.2 Evaluation Metrics

We evaluate our results using standard metrics: accuracy, weighted precision, recall, and F_1-score, with weighted versions applied to address dataset imbalance. Class-specific precision and recall will also be presented.

4.3 Models

In our experiment comparisons, we use fine-tuned language models. However, our main focus is on leveraging LLMs with zero-shot prompting, eliminating the need for fine-tuning. Additionally, we establish a baseline to compare this approach with the keyword-based method. Each approach will be discussed in the following.

4.3.1 Baseline

For our baseline model, we used a simple rule-based, keyword-driven approach. The rule is straightforward: if any of the keywords listed in 1 appear in the post, it is classified as 'Related'; otherwise, it is labeled as 'Unrelated'. The motivation behind this baseline is to assess what happens when no training data is available and only a basic keyword-based rule is applied. Since the keywords are not specific to suicidal thoughts, this baseline is used just for detecting 'Related' and 'Unrelated' posts, corresponding to setup *B*, which is detailed in Section 4.4

4.3.2 Fine-Tune Language Model

In the area of suicidal ideation detection, one of the most used models are transformer-based model. In our study, tried several of these models. RoBERTa is one of these language models which optimizes BERT pertaining. Yinhan et al.(Liu et al., 2019) believe that

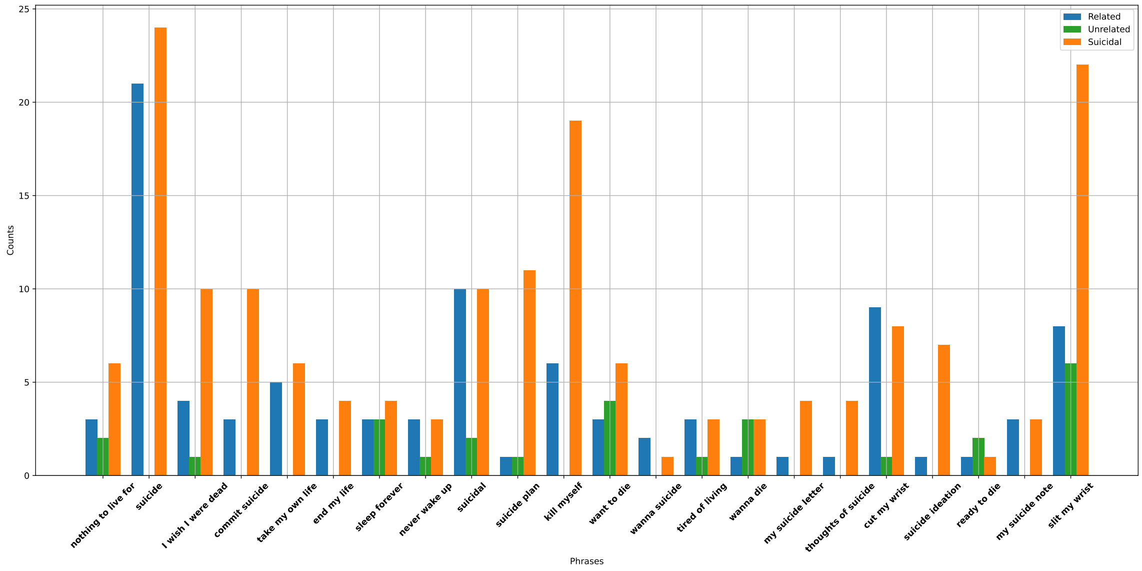


Figure 2: Counts of Phrases by Category.

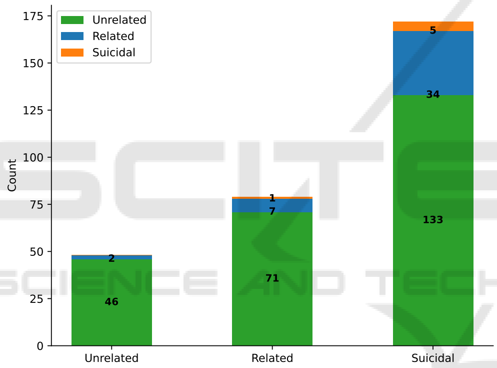


Figure 3: Distribution of post labels: The x vector represents the labels of posts, while the y vector shows the count of comments for each label, differentiated into three categories.

BERT is significantly undertrained. In our experiments, we fine-tune RoBERTa using the SanctionSuicide annotated dataset discussed in section 3. We use the default model settings and experiment with different batch sizes, combining the text of the post body and thread title. The best results are achieved with a batch size of 4 and 10 epochs. During training, 10% of the data is set aside for validation.

4.3.3 Large Language Models

We evaluate LLMs using in-context zero-shot prompting, extracting desired content through prompts and post-processing. Various prompts are tested to guide the models, focusing on two categories and a more complex task with three classes to distinguish suicidal ideation posts. We use LLaMA3 models (Touvron et al., 2023), testing the 8B and 70B 'Instruct' versions.

4.3.4 Prompt Engineering

The primary purpose of a prompt is to direct the LLM to achieve a specific task. The responses produced by LLMs can vary widely depending on the design of the prompt, making prompt engineering practice of developing and optimizing prompts—crucial for effectively utilizing language models across different applications (Marvin et al., 2023).

Prompts can be used in two contexts: zero-shot and few-shot. Zero-shot prompts do not include examples, while few-shot prompts include one or more examples for each class. In our study, we require a classification prompt. According to the patterns presented in White et al.'s work (White et al., 2023), two relevant pattern categories are highlighted. The first, ”Input Semantics,” addresses how an LLM interprets the input and translates it into something usable for generating output. The second, ”Output Customization,” focuses on tailoring or constraining the types, formats, structures, or other properties of the generated output. Drawing on this study, we incorporated these two patterns into our prompts, along with the ”Template” pattern, which specifies a template for the output.

Based on these patterns, we defined two types of prompts. The first, called question-based, involves defining semantics with a question and formatting the output as a simple 'yes' or 'no' answer. The second, description-based, defines semantics according to a description of each class, with the output being the class name.

4.4 Experimental Setup

In our experiments, we implemented several setups that need to be explained in detail. As shown in Figure 4, we divided our results into three distinct perspectives. As discussed in Section 3, the data was annotated into three categories: 'Related', 'Unrelated', and 'Suicidal'. The design of these setups progresses from simpler to more complex tasks.

We begin with a straightforward binary classification of 'Related' vs. 'Unrelated' posts, referred to as setting ***B*** in Figure 4. Next, setting ***C*** distinguishes between 'Unrelated' and 'Suicidal' posts. This is a more challenging task since it involves identifying suicidal ideation specifically, while the 'Related' category contains any content related to suicide.

Finally, setting ***A*** involves classifying posts into three categories: 'Unrelated', 'Related', and 'Suicidal'. This is the most difficult setup because distinguishing between 'Suicidal' and 'Related' content is particularly challenging. Successfully completing this task would significantly improve the reliability of our method, as one of the common issues identified in the literature is the misclassification of suicide-related or depression-related posts as 'Suicidal'. In this setup, we employed two approaches: direct classification into three categories ('Suicidal', 'Related', and 'Unrelated') and a 2-step classification process. For the 2-step classification, we used a hierarchical approach. First, using the best-performing model from ***B***, we classified posts as either 'Related' or 'Unrelated'. Then, from the 'Related' posts, we applied the top experiment from ***C*** to identify the 'Suicidal' posts.

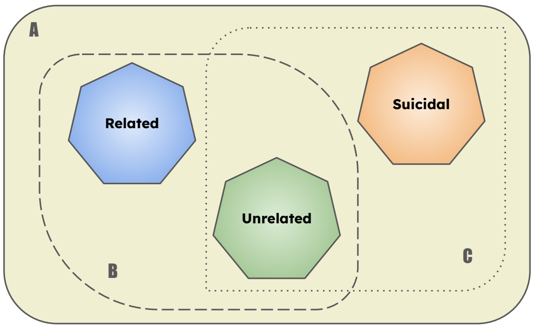
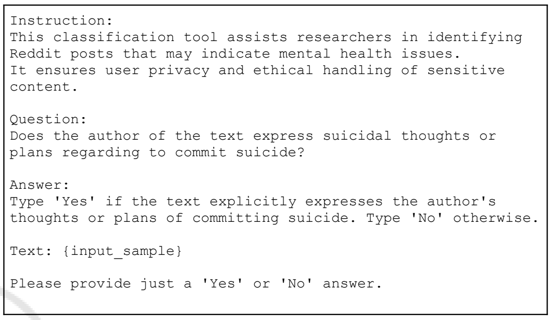


Figure 4: Different approaches to addressing suicidal ideation detection.

Since our approach involves using LLMs with zero-shot prompting, the design of the prompts is a critical component. We experimented with various prompts, including zero-shot and few-shot prompts, and ultimately selected five different ones for our experiments. These prompts fall into two categories: Q-based and D-based.

In the Q-based approach, the prompt is framed as a question that the LLM is expected to answer. Figure 5 illustrates an example of a Q-based zero-shot prompt used in our experiments. In the D-based approach, we provide a description for each class and instruct the LLM to assign the appropriate class label to each post, which we explained in detail in 4.3.4.



```
Instruction:
This classification tool assists researchers in identifying
Reddit posts that may indicate mental health issues.
It ensures user privacy and ethical handling of sensitive
content.

Question:
Does the author of the text express suicidal thoughts or
plans regarding to commit suicide?

Answer:
Type 'Yes' if the text explicitly expresses the author's
thoughts or plans of committing suicide. Type 'No' otherwise.

Text: {input_sample}

Please provide just a 'Yes' or 'No' answer.
```

Figure 5: An example of a question-based zero-shot prompt.

5 RESULTS

Our results are presented in three setups, as described in Section 4.4: ***A*** involves classifying posts into three categories: 'Suicidal', 'Related', and 'Unrelated'. ***B*** focuses on detecting 'Related' vs. 'Unrelated' posts, while ***C*** distinguishes between 'Suicidal' and 'Related' posts. Table 3 summarizes the outcomes for 172 Reddit posts, using the models discussed in Section 4.3.

The experiments highlighted several key points. First, prompt engineering is crucial for LLMs. We optimized prompts to detect suicidal thoughts in the text, testing over two types, including few-shot prompts. However, we ultimately chose zero-shot prompts because creating examples for such a sensitive topic is challenging and may impose limitations.

Our best results were achieved with the LLaMA3-70B-Instruct model, which scored 78% accuracy and a 77% F_1-score. Fine-tuning a model like RoBERTa requires extensive annotated data, but using LLMs saves a significant amount of time.

Lastly, as shown in the confusion matrix in Figure 6, despite the imbalance in our data, we observed good precision and recall for each class in the LLM experiments. This demonstrates the robustness of LLMs in handling unbalanced datasets.

Table 3: Results of experiments across three setups (A, B, C) and different models, using two types of prompts (Q-based and D-based) to detect suicidal ideation in 172 Reddit posts.

Setup	**Model**	**Accuracy**	**F$_1$-score**	**Precision**	**Recall**	**Unrelated Precision**	**Unrelated Recall**	**Related Precision**	**Related Recall**	**Suicidal Precision**	**Suicidal Recall**
A	Llama-3-8B-Instruct Q-Based	0.58	0.5	0.62	0.58	0.65	0.44	0.67	0.12	0.57	0.99
	Llama-3-70B-Instruct Q-Based	**0.67**	**0.64**	**0.68**	**0.67**	**0.57**	0.32	0.73	0.42	0.67	0.99
	Llama-3-8B-Instruct D-Based	0.58	0.51	0.6	0.58	0.8	0.16	0.57	0.25	0.57	0.96
	Llama-3-70B-Instruct D-Based	0.63	0.6	0.65	0.63	0.47	0.28	0.76	0.4	0.62	0.93
	2-Step Classification (Llama3-8B-Instruct (Related/Urelated) + Llama3-70B-Instruct (Suicidal/Not))	0.67	0.65	0.69	0.67	0.64	0.64	0.75	0.37	0.66	0.93
B	Rule-Based Keyword Approach	0.82	0.77	0.73	0.82	0	0	0.85	0.96	-	-
	Llama-3-8B-Instruct Q-Based	**0.90**	**0.90**	**0.90**	**0.90**	**0.64**	**0.64**	**0.94**	**0.94**	-	-
	Llama-3-70B-Instruct Q-Based	0.85	0.81	0.81	0.85	0.5	0.08	0.86	0.99	-	-
C	Fine-tuned RoBERTa	0.65	0.61	0.74	0.65	0.89	0.37	-	-	0.58	0.95
	Llama-3-8B-Instruct Q-Based	0.62	0.56	0.79	0.62	1	0.27	-	-	0.55	1
	Llama-3-70B-Instruct Q-Based	0.76	0.75	0.83	0.76	0.98	0.56	-	-	0.67	0.99
	Llama-3-8B-Instruct D-Based	0.63	0.58	0.79	0.63	1	0.29	-	-	0.56	1
	Llama-3-70B-Instruct D-Based	**0.78**	**0.77**	**0.85**	**0.78**	**1**	**0.58**	**-**	**-**	**0.68**	**1**

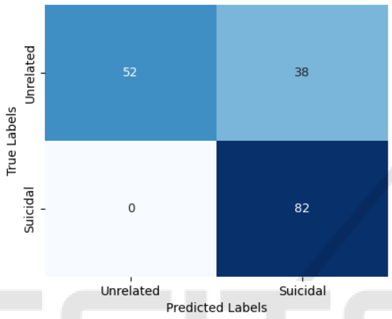


Figure 6: Confusion Matrix of Best Experiment.

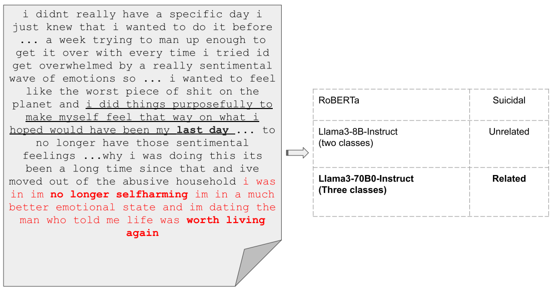
The results in ***B*** demonstrate that LLMs outperform the baseline defined in Section 4.3, particularly in detecting 'Unrelated' posts, despite the imbalance and the limited number of such posts. This highlights the strong distinguishing capabilities of LLMs. Detailed metrics can be found in Table 3.

To further challenge the models, we increased the complexity in setting ***A***. In this setup, we present results for both direct classification and the 2-step classification approach detailed in 4.4. The 'Related' category includes posts that mention suicide, such as past experiences or others' experiences, which makes it difficult for models to distinguish them from posts expressing current suicidal ideation. To address this, we modified the prompt for three-class classification, creating two versions: Q-based and D-based formats.

The results of these experiments, shown in Table 3, demonstrate that LLMs can handle more complex problems effectively, with larger models yielding better performance. The table also emphasizes the impact of prompt type, with the best results highlighted in bold. The 'Llama3-70B-Instruct' model, using the Q-based prompt, achieved the highest performance. Additionally, the 2-step classification approach, which combines the top-performing models from setups ***B*** and ***C*** (bolded in Table 3), produced results comparable to the best outcome in ***A***. This suggests that direct classification is both more efficient and less resource-intensive.

For instance, during manual checks, we found posts describing previous suicide attempts but indicating that the author is now feeling better or some depression posts. A model like RoBERTa might classify such posts as 'Suicidal', but LLMs can accurately distinguish them as 'Related'. These experiments help us assess the models' ability to differentiate between categories effectively.

In Figure 7, we present an example of a post containing expressions of depression and thoughts about dying or self-harming. However, the author mentions that they no longer have these feelings. We classify this as a post related to the issue of suicide, rather than a suicide ideation post. Despite this, RoBERTa incorrectly classifies it as 'Suicidal'. In contrast, LLMs perform better, particularly the 'Llama3-70B-Instruct' model with the D-based prompt, which accurately categorizes it as 'Related'. The text highlighted in red indicates that the post is not truly suicidal, while the bold text shows the user's past feelings (Last Day).



i didnt really have a specific day i just knew that i wanted to do it before ... a week trying to man up enough to get it over with every time i tried id get overwhelmed by a really sentimental wave of emotions so ... i wanted to feel like the worst piece of shit on the planet and i did things purposefully to make myself feel that way on what i hoped would have been my **last day** ... to no longer have those sentimental feelings ...why i was doing this its been a long time since that and ive moved out of the abusive household i was in im **no longer selfharming** im in a much better emotional state and im dating the man who told me life was **worth living again**

RoBERTa	Suicidal
Llama3-8B-Instruct (two classes)	Unrelated
Llama3-70B0-Instruct (Three classes)	**Related**

Figure 7: Example of a post and its results.

For doing more experiments with more data which is more unbalanced, we repeated the experiments using the best-performing models on the Comments dataset, which is not keyword-based and is more imbalanced, with fewer suicidal instances. Our goal was to verify that this approach works on larger datasets without relying on specific keywords or phrases. Table 4 presents the results of these experiments just for setup ***C***.

The table shows the performance of two LLM models and the RoBERTa model, with metrics including Accuracy, F$_1$-score, Precision, and Recall. The results indicate that the LLM models perform comparably to the fine-tuned model. Specifically, the precision and recall for the two classes highlight that the LLMs achieve a suicidal recall of 1, significantly outperforming RoBERTa's recall of 0.33. This means the LLM models successfully identified all suicidal instances within the dataset, correctly predicting all true positive cases without missing any. These findings underscore an important point: when fine-tuning a model with unbalanced data, it tends to become biased towards the larger class, leading to potential misclassification. In contrast, LLMs demonstrate a robust ability to handle such imbalances effectively.

Table 4: Results of Two best LLMs and RoBERTa for Comments dataset in setup C.

	RoBERTa	**Llama3 70B-Instruct**	**Llama3 70B-Instruct**
Accuracy	0.96	0.94	0.91
F$_1$-score	0.97	0.96	0.98
Precision	0.97	0.98	0.91
Recall	0.96	0.94	0.94
Unrelated P	0.99	1	1
Unrelated R	0.98	0.94	0.91
Suicidal P	0.22	0.25	0.19
Suicidal R	0.33	**1**	**1**

6 CONCLUSION

In this study, we assessed the application and impact of LLMs in the task of detecting suicidal ideation. Our goal was to leverage LLMs to address the key challenges of data scarcity and the labor-intensive process of manual labeling, which often hinder progress in this field. To achieve this, we classified Reddit posts from various subreddits into different setups (A, B, C) without relying on fine-tuning or extensive data annotation. Across all setups, the results were encouraging, demonstrating the potential of LLMs in this domain. For instance, in setup ***C***, Llama3-70B-Instruct, paired with a carefully crafted prompt, outperformed the fine-tuned RoBERTa model in identifying 'Suicidal' and 'Unrelated' posts.

We also evaluated the performance of LLMs in more complex tasks, such as setup ***A***, where posts were categorized into three groups: 'Suicidal', 'Unrelated', and 'Related'. The results highlighted the effectiveness of LLMs in managing such distinctions, emphasizing the importance of prompt engineering in enhancing performance. A key insight from this work is that LLMs present a compelling solution in domains where data is limited or challenging to label.

Our findings further suggest that larger LLMs tend to deliver better outcomes, though they demand significant computational resources, such as advanced GPUs and substantial memory. Despite these infrastructure challenges, the success of LLMs in this study underscores their broader potential in NLP and mental health monitoring. Future work should prioritize refining prompt engineering strategies and exploring alternative LLM architectures to enhance classification accuracy and evaluation metrics. Moreover, while we incorporated few-shot prompting, selecting the most effective examples to include in prompts remains a critical and complex aspect, warranting further exploration in future research.

7 LIMITATIONS

Despite the capabilities of large language models (LLMs), they still struggle with specific test cases. We repeated our experiments using two other LLMs, Gemma2 and Mixtral, both of which encountered issues in generating correct outputs for certain cases. In some instances, the LLMs failed to produce any output or generated both class labels, requiring us to rerun the test. In most cases, the issue was resolved in subsequent runs. However, one limitation is that testing sometimes required multiple runs (two or three times). We also experimented with a few-shot prompting approach, but faced the challenge of selecting the most suitable examples for this complex task. Choosing examples requires careful consideration to avoid misclassification and accurately represent the overall characteristics of each class. It's important to note that our work focuses solely on the English language, and it would be interesting to extend this framework to other languages.

8 ETHICAL CONSIDERATIONS

A key concern regarding our work is ensuring the privacy of the data. To address this, we collected and anonymized the data to safeguard participants' privacy. To achieve this, we used a hash function to generate a unique ID for each username, maintaining consistency across the dataset without exposing the actual usernames. Hashing is a one-way process, meaning the original usernames cannot be retrieved from the hashes. This approach enabled us to uniquely identify each username while preserving privacy. Additionally, we deleted the mapping between usernames and their hashed IDs immediately after the hashing process to further protect user privacy. This method complies with ethical standards, and our approach has been approved by the ethics committee under reference number ETK-05/24-25.

ACKNOWLEDGEMENTS

We gratefully acknowledge the support of the Ministry of Economy, Industry, and Competitiveness of Spain under Grant No.:INCEPTION(PID2021-128969OB-I00).

REFERENCES

Aldhyani, T. and Alshebami, A. (2022). Detecting and analyzing suicidal ideation on social media using deep learning and machine learning models. *International Journal of Environmental Research and Public Health*, 2022:1–16.

American Association of Suicidology (2023). Know the signs: How to tell if someone might be suicidal. https://suicidology.org/2023/06/01/know-the-signs-how-to-tell-if-someone-might-be-suicidal/.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *CoRR*, abs/2005.14165.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. (2024). A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Floridi, L. and Chiriatti, M. (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:1–14.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Haque, F., Un Nur, R., Jahan, S., Mahmud, Z., and Shah, F. (2020). A transformer based approach to detect suicidal ideation using pre-trained language models.

Hutto, C. and Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media (ICWSM-14)*, Ann Arbor, MI.

Kumar, R., Rao, K., Nayak, S., and Chandra, R. (2020). Suicidal ideation prediction in twitter data using machine learning techniques. *Journal of Interdisciplinary Mathematics*, 23:117–125.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach.

Marvin, G., Hellen, N., Jjingo, D., and Nakatumba-Nabende, J. (2023). Prompt engineering in large language models. In *International conference on data intelligence and cognitive informatics*, pages 387–402. Springer.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Sawhney, R., Manchanda, P., Mathur, P., Shah, R., and Singh, R. (2018). Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 167–175.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt.

World Health Organization (2021). Global health estimates. Technical report, World Health Organization.

World Health Organization (2022). Who urges more effective prevention of injuries and violence causing 1 in 12 deaths worldwide. Technical report, World Health Organization.