Leveraging LLMs for Translating and Classifying Mental Health Data

Konstantinos Skianis[*], **A. Seza Doğruöz**[#], **John Pavlopoulos**[†¶]
[*] Department of Computer Science and Engineering, University of Ioannina, Greece
[#] LT3, IDLab, Universiteit Gent, Belgium
[†] Department of Informatics, Athens University of Economics and Business, Greece
[¶] Archimedes/AthenaRC, Greece
kskianis@cse.uoi.gr as.dogruoz@ugent.be annis@aueb.gr

Abstract

Large language models (LLMs) are increasingly used in medical fields. In mental health support, the early identification of linguistic markers associated with mental health conditions can provide valuable support to mental health professionals, and reduce long waiting times for patients. Despite the benefits of LLMs for mental health support, there is limited research on their application in mental health systems for languages other than English. Our study addresses this gap by focusing on the detection of depression severity in Greek through user-generated posts which are automatically translated from English. Our results show that GPT3.5-turbo is not very successful in identifying the severity of depression in English, and it has a varying performance in Greek as well. Our study underscores the necessity for further research, especially in languages with less resources. Also, careful implementation is necessary to ensure that LLMs are used effectively in mental health platforms, and human supervision remains crucial to avoid misdiagnosis.

1 Introduction

Mental health issues (e.g., depression, anxiety, and post-traumatic stress disorder (PTSD)) are prevalent worldwide and pose significant challenges to public health (World Health Organization, 2021). Traditional methods for diagnosing mental health conditions often rely on self-reported surveys, clinical interviews, and standardised assessments conducted by trained professionals (Kessler and Üstün, 2004). While these methods are effective, they are also resource-intensive, time-consuming, and may not always be accessible to individuals in need, particularly for speakers of languages beyond English.

In this context, the application of LLMs to detect mental health symptoms from textual data offers a compelling alternative. These models can analyse large volumes of text data (e.g., social media posts, forum discussions, and personal narratives) quickly to identify linguistic markers associated with mental health conditions (Guntuku et al., 2019; Chancellor et al., 2019). This capability opens up new avenues for early detection and intervention, providing valuable support to mental health professionals and potentially reaching out to the patients whose symptoms may be overlooked and/or save time (e.g., long waiting times).

Despite the potential benefits, the performance of LLMs in multilingual mental health symptom detection remains underexplored. Previous studies have primarily focused on English-language datasets, leaving a gap in our understanding of how these models perform in other linguistic contexts (Raihan et al., 2024). Hence, our work raises the following research questions:

- Can an LLM accurately predict the severity of mental health conditions from English user-generated posts?
- Is the detection performance similar if one automatically translates the English posts to another language (e.g., Greek) with LLMs?

To address these research questions, first, we assess a state-of-the-art multilingual LLM when predicting the severity of mental health in English user-generated posts. Then, we automatically translate these posts from English to Greek, a language for which there are no resources for this task (Bakagianni et al., 2025), and re-assess the performance of the LLM. Our research not only contributes to the development of more robust and inclusive AI-driven mental health diagnostic tools but also emphasises the importance of culturally and linguistically sensitive approaches in mental health care beyond English. The contribution of this work lies into the evaluation of the predictive power of a popular LLM in detecting the severity of depression across English and Greek.

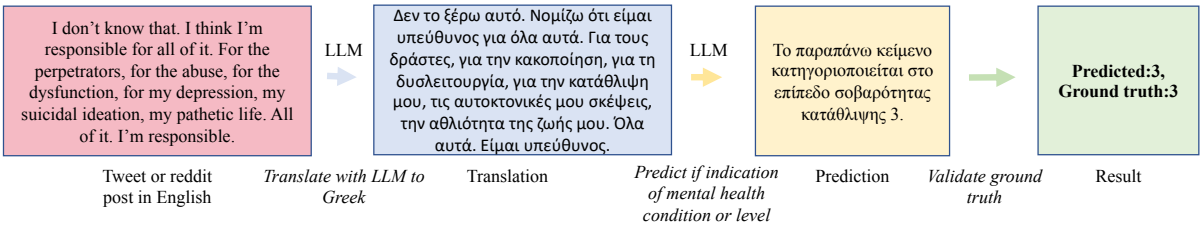

Figure 1: An illustration of our proposed methodology.

2 Related work

LLMs have remarkable accuracy in detecting mental health symptoms by leveraging their ability to understand context and semantics at a deeper level. Examples include BioBERT (Lee et al., 2020), and ClinicalBERT (Huang et al., 2019), which are pre-trained on biomedical corpora or clinical notes. In contrast, models like MentalBERT (Ji et al., 2022), and DisorBERT (Aragón et al., 2023) are pre-trained on mental health-related social media data. Additionally, research by Benton et al. (2017) showed that NLP can effectively assess depression and PTSD from clinical notes, further validating the utility of these models in a healthcare setting.

Although the details about training and evaluation are not always transparent, the multilingual capabilities of LLMs enable these models to understand and generate text in various languages. A recent example is the XLM-R model, which has been trained on a vast amount of multilingual data and shows strong performance across multiple languages. According to Conneau et al. (2020), their model XLM-R outperforms previous models on a wide range of tasks, demonstrating that leveraging large-scale multilingual data can lead to improvements in cross-lingual understanding.

Despite these advancements, significant challenges remain in achieving truly equitable performance across all languages and handling culturally specific contexts accurately (Zhang et al., 2020). Languages with limited digital text data still pose a considerable challenge for LLMs, often resulting in lower performance and less reliable outputs. Addressing this issue requires more inclusive data collection practices and further research into transfer learning techniques that can better utilise limited resources (Doğruöz and Sitaram, 2022). Additionally, capturing cultural nuances and context-specific meanings is a complex task, as language is deeply intertwined with cultural and societal norms. Efforts to improve these aspects include developing more sophisticated algorithms and incorporating diverse and representative datasets (Doğruöz et al., 2023), ensuring that the benefits of multilingual LLMs are accessible to a broader range of users globally.

More recently, a plethora of social network datasets targeting mental health, have been available (Raihan et al., 2024). The authors gathered social media posts from Reddit and Twitter regarding depression, PTSD, schizophrenia, and eating disorders. Moreover, multiple models were fine-tuned on small-sized publicly available annotated mental health datasets by the authors to use them for labelling their introduced MentalHelp dataset. Nevertheless, the dataset includes only posts in English, and thus its use is restrictive, disallowing further research for multilingual scenarios.

3 Proposed Methodology

Our methodology leverages LLMs, in order to translate English social media posts to another language (Greek), and then to predict mental health conditions accordingly. Specifically, we translate the social media posts to Greek via an LLM, and we feed the resulting translations to a prompt that asks the LLM to predict specific severity levels of mental health conditions. We assess the LLM by comparing the predicted classes in both languages against the ground truth labels. We note that although our study focuses on Greek, our method is applicable to other language pairs as well. An illustration of the proposed approach for evaluating LLMs for multilingual detection of mental health conditions is shown in Figure 1.

4 Experiments

We select the DEPSEVERITY dataset of Naseem et al. (2022), which consists of posts from the social media platform Reddit, regarding different levels of depression. The posts (in English) are already labelled in terms of four levels of severity: minimal,

Dataset	Category	#Classes	#Instances	Labels (#Support)	Prompt
DEPSEVERITY Naseem et al. (2022)	Depression	4	3553	Minimum (2587) Mild (290) Moderate (394) Severe (282)	"Categorise the following text with 1 of the 4 depression severity levels (0: Minimum, 1: Mild, 2: Moderate, 3: Severe)"

Table 1: The benchmark dataset used in our study along with statistics.

	English			Greek		
Class	**Pr**	**Rec**	**F1**	**Pr**	**Rec**	**F1**
MINIMUM	0.98	0.14	**0.25**	0.99	0.07	0.14
MILD	0.04	0.15	**0.07**	0.04	0.17	0.06
MODERATE	0.13	0.22	0.17	0.14	0.55	**0.23**
SEVERE	0.13	0.71	**0.22**	0.16	0.28	0.20
Macro avg	0.32	0.30	**0.17**	0.33	0.27	0.16

Table 2: **GPT-3.5 with 0-shot learning on DEPSEVERITY**, measuring Precision, Recall, and F1 per class in English and Greek. The last row shows the macro averages. The best F1 per class is shown in bold.

mild, moderate, and severe depression. The majority of posts belong to the minimal severity level (Table 1) making it a highly imbalanced dataset. We specifically selected this multi-class dataset to make the task more challenging for the LLM, as binary problems would have been easier to answer.

We use GPT3.5-turbo (Brown et al., 2020) through its API to translate the posts and predict the labels. The temperature parameter is set to 0, so the outcome is reproducible, regarding translations and predictions. We approach the task with text classification, comparing the predicted classes with the ground-truth ones, reporting Precision, Recall and F1. We experiment with English as the source and Greek as the target language. The prompt we used to predict the severity levels is shown in Table 1.

Preliminary Prompting Before exposing our LLM to any posts, definitions, or instructions, either for the translation or the classification task, we asked how it would classify posts to different levels of depression severity. The response of LLM was that it would initially try to identify language patterns associated with depression, such as:

- Persistent negative emotions, such as sadness, or hopelessness.
- Self-criticism or feelings of worthlessness.
- Expressions of loneliness or social withdrawal.
- Changes in behavior or routines, as in sleep patterns or appetite.
- References to emotional pain or distress.

More specifically, it would try to adapt the four depression severity levels to fit the context of social media posts, as follows.

- Level 1 (Minimum): Posts with minimal or occasional expressions of sadness.
- Level 2 (Mild): Posts indicating frequent negative emotions or noticeable changes in behavior.
- Level 3 (Moderate): Posts suggesting significant impairment in daily functioning or clear signs of distress.
- Level 4 (Severe): Posts indicating severe emotional distress, potential risk factors for self-harm, or complete social withdrawal.

We can infer that the LLM expects posts with very generic indications of negative signs.

Classification in the source language Initially, we experimented with the data in their source language (English), to set the baseline performance. That is, no translation step has been performed at this stage. As we observe in Table 2, the best F1 is achieved for the lowest severity/indication (F1=0.25) and the next best for the highest severity (F1=0.22). The overall low performance (F1=0.17) can be attributed to the difficulty of the task of detecting specific levels of depression, which are considered less distinct compared to other conditions. Therefore, it is likely more challenging for an LLM to distinguish these levels in user posts.

Classification in the target language In Greek, the worst results are observed for the mild level (F1=0.06), similarly to English (F1=0.07). Overall, a drop in scores is observed across all classes except for the moderate level, where results improve (F1=0.23; from 0.17). Also, although the scores for the two edges remain relatively high, the score for the lowest severity dropped in Greek (F1=0.14).

Error Analysis Mental health terminologies and nuances may not be well-represented in the available Greek corpora, making it difficult for an LLM

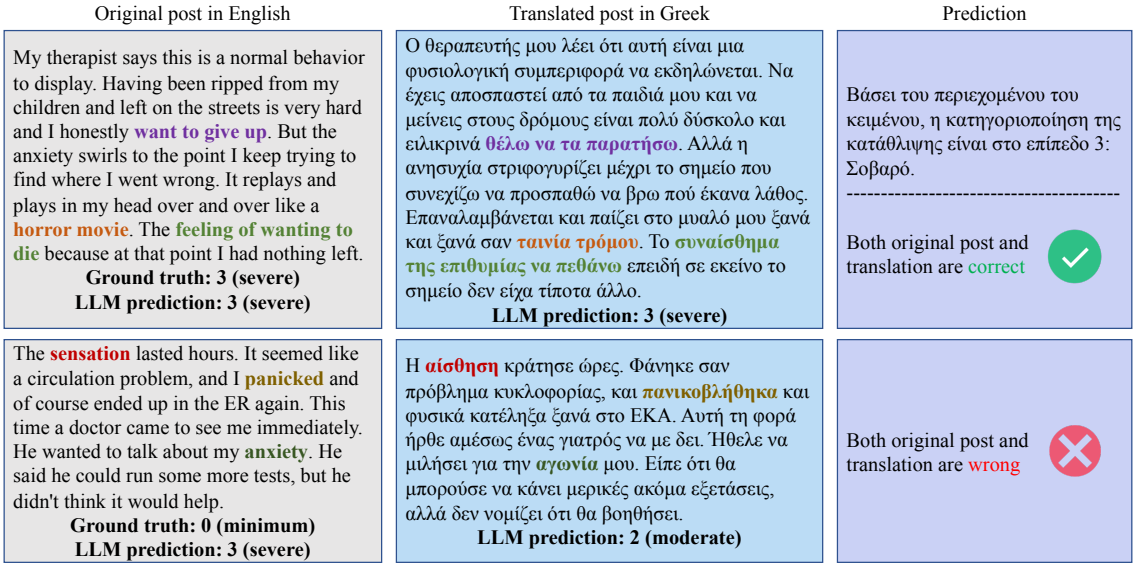

Figure 2: Example translation (from English to Greek), with similar colour used for original and translated words.

to grasp the context accurately. Figure 2 presents two instances of the dataset and their corresponding translations in Greek. We marked words and their translations with similar colours for better visibility. Both translations appear to be accurate and convey the same meaning as the original Greek text. There are no significant differences that would alter the understanding of the texts. The first segment contains explicit mentions of severe depression symptoms such as “want to give up" and “feeling of wanting to die.” These statements clearly indicate a severe level of depression, which is why both the ground truth and prediction were classified as severe. The second segment describes physical sensations, panic, and anxiety but does not express a severe depressive state. The ground truth classified this as minimal depression, likely because the primary issues are related to panic and anxiety rather than depression. The LLM predicted a moderate level of depression for the second segment, possibly because it picked up on the words “panicked" and “anxiety", which are associated with higher levels of distress. However, these symptoms are more indicative of anxiety disorders rather than depression. The discrepancy in the second prediction can be attributed to the LLM’s interpretation of anxiety and panic as indicative of moderate depression, whereas the ground truth assessment considers these symptoms in the context of a panic or anxiety disorder with minimal depression.

Cost of experiments The total cost of credits using the GPT3.5-turbo API was less than $30 (US dollars), showing that minimal resources were required to conduct our experiments, without the need for expensive GPU infrastructure or fine-tuning. Our cost-saving methodology for utilizing resources efficiently is especially promising for extending medical data sets in English into other languages.

5 Conclusion

In our study, we focused on the ability of an LLM to predict the severity of depression in user-generated posts in English (source language) and in Greek (target language) when the posts are machine-translated by the same LLM. Our findings show that there is room for improvement in the source language (English) and that the edge classes are easier to handle. In the target language (Greek), results dropped for all but the moderate level, for which results increased considerably. Considering the varying performance of the LLM across the two languages, there is a need for utmost precautions not to rely on LLMs solely for translation in any healthcare setting. As stated by Stade et al. (2024), diagnosis of mental health should never be left alone to automatic systems, and it should never replace the diagnosis by human professionals, to avoid possible errors and/or misdiagnoses. Our approach, however, does not aim to assist the patients. By contrast, it is potentially useful to *train* professionals in the mental healthcare domain, which can be vital for languages other than English.

Acknowledgments

In this project, John Pavlopoulos was partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. We thank OpenAI for granting us free credits for research purposes.

Limitations

Translation In this work, we used a popular LLM like GPT3.5 to translate posts. Translating using only an LLM and not having an expert or native-language human resources introduces a small loss of information that in some cases affects the final results.

Evaluation Automatically evaluating the performance of LLMs is by definition a hard task. In order to measure the performance we search for the label in the LLM output. Whenever no label is detected we count it as the minimum label (class: 0) for the depression dataset and not suicidal (class: 0) for the suicide dataset.

Potential risks The quality of publicly available datasets, especially in sensitive areas like the mental health care domain is of great importance for prediction tasks. The data sets we employed as a basis in our study, along with our created multilingual data should be used with utmost care and only for assisting the health care specialists instead of diagnosing patients directly.

References

Mario Ezra Aragón, A Pastor López-Monroy, Luis C González, David E Losada, and Manuel Montes. 2023. Disorbert: A double domain adaptation model for detecting signs of mental disorders in social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15305–15318.

Juli Bakagianni, Kanella Pouli, Maria Gavriilidou, and John Pavlopoulos. 2025. A systematic survey of natural language processing for the greek language. *Patterns*, 6(11).

Adrian Benton, Margaret Mitchell, Dirk Hovy, et al. 2017. Multitask learning for mental health conditions with limited social media data. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017. Proceedings of Conference*. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 79–88.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

A. Seza Doğruöz and Sunayana Sitaram. 2022. Language technologies for low resource languages: Sociolinguistic and multilingual insights. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 92–97, Marseille, France. European Language Resources Association.

A. Seza Doğruöz, Sunayana Sitaram, and Zheng Xin Yong. 2023. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5751–5767, Singapore. Association for Computational Linguistics.

Sharath Chandra Guntuku, Raphael Schneider, Ashley Pallavi Pelullo, Jennifer Young, Vivienne Wong, Lyle H Ungar, and Daniel E Polsky. 2019. Studying expressions of loneliness in individuals using twitter: An observational study. *BMJ Open*, 9(11):e030355.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv:1904.05342*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. Mentalbert: Publicly available pretrained language models for mental healthcare. In *proceedings of the thirteenth language resources and evaluation conference*, pages 7184–7190.

Ronald C Kessler and T Bedirhan Üstün. 2004. The world mental health (wmh) survey initiative version of the world health organization (who) composite international diagnostic interview (cidi). *International Journal of Methods in Psychiatric Research*, 13(2):93–121.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Jungyun Seo, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language

representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Usman Naseem, Adam G Dunn, Jinman Kim, and Matloob Khushi. 2022. Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022*, pages 2563–2572.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Mentalhelp: A multi-task dataset for mental health in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11196–11203.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.

World Health Organization. 2021. Mental health. Retrieved from https://www.who.int/health-topics/mental-health.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.