Depression detection in Modern Greek

Vivian Stamou[1], **George Mikros**[2], **George Markopoulos**[1], **Spyridoula Varlokosta**[1]

[1]National and Kapodistrian University of Athens, Greece,
[2]Hamad Bin Khalifa University, Qatar, Affiliation3
Panepistimiopoli, Zografou 157 72
Education City, Doha, Qatar 34110
vivianstamou@gmail.com, gmikros@hbku.edu.qa,
{gmarkop, svarlokosta}@phil.uoa.gr

Abstract

Despite advancements in NLP-based mental health screening, research remains predominantly English-centric, leaving under-resourced languages insufficiently explored. This study investigates depression detection in Modern Greek social media through a series of experiments. We benchmark traditional machine learning (ML) models against transformer architectures (GreekBERT, GreekSocialBERT, mBERT, and XLM-R) under two settings: a topic-oriented control corpus and a high-similarity stress-test contrasting a gold case of a depressed user with a matched control. Transformer models consistently outperform ML models (F1 = 0.95) but offer limited interpretability. To address this limitation, we incorporate LIWC-derived psycholinguistic features with SHAP explanations to examine model behavior in relation to established linguistic markers. The analysis reveals linguistic patterns consistent with depressive symptoms, such as reduced work-related engagement, social withdrawal, and the motivational deficits characteristically linked to anhedonia in clinical literature. Overall, the results provide a baseline for depression detection in Modern Greek and underscore the importance of grounding automated screening in clinically interpretable evidence.

Keywords: depression detection, social media, mental health

1. Introduction

Natural Language Processing (NLP) research plays a pivotal role in advancing mental health disorders such as depression screening through the development of sophisticated speech and text based models (Gómez-Zaragozá et al., 2025; Liu et al., 2022). To date, research has predominantly utilized social media data (De Choudhury et al., 2013; Eichstaedt et al., 2018), with a heavy reliance on platforms like X (formerly Twitter) and Reddit (Harrigian et al., 2021). While the majority of these studies focus on English-language resources, recent efforts have begun to bridge the gap for other languages, including Portuguese (Santos et al., 2023), German (Zanwar et al., 2023), Arabic (Almouzini et al., 2019), and Chinese (Zhang et al., 2024).

This shift toward linguistic diversity is exemplified by recent efforts to consolidate multilingual research, such as the first comprehensive survey of mental disorder detection in non-English languages (Bucur et al., 2025). Exploring diverse languages is essential to determine whether linguistic cues of depression are universal or subject to cultural and contextual variation.

Notwithstanding this progress, two significant gaps remain in the literature: (i) methodological opacity; studies utilize Machine Learning (ML) or Deep Learning (DL) architectures to classify depressed vs. non-depressed individuals (Tadesse et al., 2019; Husseini Orabi et al., 2018) which lack interpretability required for clinical confidence and (ii) lack of explainability; existing methods fail to provide the reasoning behind a classification. In mental health contexts, it is crucial to understand how automated markers correlate with established clinical symptomatology (Zhang et al., 2022).

In this work, we present a comprehensive evaluation of text-based models using a social media dataset in Modern Greek (MG). Our study establishes a performance baseline for depression screening in an under-resourced language, contributing to the broader goal of linguistic inclusivity in NLP. Furthermore, we leverage Linguistic Inquiry and Word Count (LIWC) features (Pennebaker et al., 2015) to enhance model interpretability. By mapping specific linguistic patterns to classification outcomes, we investigate the extent to which these cues align with documented clinical symptoms of depression.

The remainder of this paper is organized as follows: Section 2 provides a detailed description of the Modern Greek datasets utilized in this study. Section 3 outlines the experimental framework, detailing the preprocessing pipeline and the selection of both traditional baselines and transformer-based models. In Section 4, we present the experimental results and analysis. Section 5 summarizes our findings and suggesting avenues for future research in multilingual mental health NLP.

Proceedings of the 6th RaPID@MENTAl.ai Workshop @ LREC 2026, pages 106–114
12 May 2026.

2. Dataset description

Previous research highlights the effectiveness of self-disclosure statements, where users are identified based on explicit mentions of a depression diagnosis, as reliable proxies for depression detection in social media corpora (Jagfeld et al., 2021; Jamil et al., 2017; Coppersmith et al., 2015). Following these established practices, our study utilizes the Modern Greek depression and control corpora compiled by Stamou et al. (2024).

In particular, the Depression Corpus (DC) was developed by isolating tweets containing explicit self-reports, specifically: 'I have been diagnosed with depression. The dataset includes 51 unique users, whose profiles underwent manual filtration to exclude instances of humor, off-topic references, and automated health news feeds. Subsequently, a comprehensive corpus of 659,189 tweets was constructed by retrieving the longitudinal tweet history for these individuals.

In addition to this, the work utilizes two distinct control datasets. The first, Control Corpus 1 (CC1), follows the framework of Chancellor and Choudhury (2020) and consists of non-depressed users identified through random sampling. This dataset was restricted to Greek-language users with no documented history of mental health disorders (i.e. no reference to mental health issues). The second, Control Corpus 2 (CC2), is a topic derived corpus to ensure that the control group reflects the same thematic concerns typically expressed by users in the depression group. By matching the control data to these specific areas of interest, the authors aimed to minimize thematic bias and ensure that classification relies on linguistic markers of depression rather than mere differences in subject matter.

This combination of control corpora provided a reliable, balanced foundation for subsequent depression detection experiments. From the initial pool of 659,189 tweets, we extracted randomly a subset of 10K posts for the modeling phase. This subsampling strategy was employed to ensure a balanced distribution across the depression and control classes, preventing model bias toward high-volume users while maintaining computational efficiency.

3. Experimental setup

3.1. Data preprocessing

The primary objective of this study was to develop a binary classifier capable of distinguishing between depressed and non-depressed individuals. To ensure high data quality before model training, the corpora were preprocessed by removing URLs and normalizing repetitive punctuation (e.g., "!!!!"). We utilized the Stanza library (Qi et al., 2020) for robust tokenization and performed stopword removal to reduce feature noise. Furthermore, we excluded all tweets containing fewer than two words to ensure sufficient linguistic context for the models.

3.2. Baseline

For the experiments, we employed PyCaret (Ali, 2020), an open-source tool that provides several ML classifiers. This allowed for the systematic benchmarking and comparison of twelve distinct machine learning classifiers, leveraging its integrated modules for model selection and performance evaluation.

- Linear & Statistical: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), SVM (Linear Kernel), and Naive Bayes (NB)
- Tree-Based: Random Forest (RF), Extra Trees (ET), AdaBoost, and Decision Trees (DT)
- Distance Based: K-Neighbors (KNN) and Multi-layer Perceptron (MLP)
- Baseline: A Dummy Classifier was included for performance comparison.

Text was encoded using two distinct feature extraction approaches:

- TF-IDF vectorization: We used an n-gram range of (1, 3) with a maximum of 5,000 features. This approach mitigates the dominance of high-frequency words by calculating the Term Frequency-Inverse Document Frequency, thereby amplifying the significance of rare, semantically rich terms.
- LIWC Lexicon: Features were extracted using the Linguistic Inquiry and Word Count (LIWC) tool. Unlike the TF-IDF approach, punctuation and accentuation were retained here, as the LIWC dictionary utilizes these markers to capture psychological and emotional states. In Greek, removing diacritics may merge distinct lexical forms, potentially affecting LIWC category assignments. Additionally, punctuation marks (e.g., exclamation and question marks) contribute to the detection of affective intensity and cognitive processes, which are central to LIWC-based analysis.

Experiments were conducted using an 80/20 train-test split with 5-fold cross-validation. Dataset splitting was performed at the user level rather than the tweet level. Specifically, tweets were grouped by unique user IDs prior to splitting, ensuring that all tweets from a given user appear exclusively in

either the training or test set. All feature variables were scaled using Z-score normalization to ensure a mean of 0 and a standard deviation of 1.

3.3. Transformer-based models

We evaluate several transformer-based models to assess the impact of different pretraining corpora on depression detection in MG:

- `bert-base-greek-uncased-v1`: A BERT model specifically trained on Modern Greek corpora.
- `greeksocialbert-base-v2`: A model pretrained on Greek social media data, potentially capturing platform-specific nuances.
- `bert-base-multilingual-cased`: A multilingual model trained on 104 languages, including Greek.
- `xlm-roberta-base`: A robust multilingual model optimized for cross-lingual transfer.

All models were fine-tuned for 4 epochs with a learning rate of 10^{-5} and a batch size of 16. The experiments were executed in a Google Colaboratory environment equipped with 12 GB of Virtual Memory and an NVIDIA Tesla T4 GPU. For the machine learning baselines described in Section 3.2, all feature variables (TF-IDF and LIWC) were scaled using Z-score normalization to ensure a mean of 0 and a standard deviation of 1. To ensure the reproducibility of our results, the session seed was fixed to `123`.

4. Results

This section presents the experimental results on depression detection and their analysis.

4.1. Baseline

Evaluation on the CC2 We conducted the experiments on a a balanced sample of the corpus, consisting of 10K tweets in total (5,000 from the depression corpus and 5,000 from the CC2 control group). As illustrated in Table 1, we observed that increasing the training set size beyond this point yielded diminishing returns. Doubling the corpus to 40K tweets resulted in only marginal gains in the F1-score, suggesting that 10K tweets provide a sufficient representation of the underlying linguistic patterns for this task.

Table 2 summarizes the performance of twelve classifiers using LIWC features. The evaluation includes standard metrics such as Accuracy, Precision, Recall, and F1-score, along with the Kappa statistic, which assesses the agreement between predicted and true labels while accounting for chance. Additionally, the Matthews Correlation Coefficient (MCC), a robust measure that considers true and false positives and negatives, is presented.

Train Size	Acc	Prec	Rec	F1	F_{mac}
10k	0.9400	0.9434	0.9367	0.9398	0.9399
20k	0.9534	0.9498	0.9575	0.9536	0.9534
40k	0.9576	0.9453	0.9716	0.9582	0.9576

Table 1: Performance as a function of training set size.

At first glance, based on the Accuracy metric, the top-performing models are LightGBM (F1 = 0.7123) and Extra Trees (F1 = 0.6779). However, a closer look reveals a small imbalance between precision and recall across these models. Beyond predictive power, a primary goal of this study is to ensure model interpretability. Tree-based ensembles are particularly suited for this, as they allow for the extraction of feature importance and decision paths. Consequently, we selected the Extra Trees Classifier ('et') as our primary model for further analysis, as it offers a superior balance between competitive performance and the transparency required for LIWC feature interpretation.

Table 2: Classification results with LIWC features on the CC2.

Model	Acc.	AUC	Rec.	Prec.	F1	κ	MCC
LightGBM	**0.7170**	**0.7938**	0.7014	0.7238	0.7123	**0.4339**	0.4343
Extra Trees	0.7128	0.7759	0.6055	**0.7706**	0.6779	0.4256	**0.4358**
Random Forest	0.7125	0.7807	0.6122	0.7652	0.6801	0.4248	0.4336
Grad. Boosting	0.7125	0.7873	0.7134	0.7117	0.7125	0.4249	0.4250
AdaBoost	0.7086	0.7792	0.7262	0.7011	**0.7134**	0.4172	0.4175
KNN	0.6555	0.7148	0.7480	0.6309	0.6844	0.3112	0.3168
Decision Tree	0.6461	0.6157	0.6693	0.6392	0.6539	0.2923	0.2927
QDA	0.6161	0.7437	0.8449	0.5794	0.6874	0.2326	0.2616
Naive Bayes	0.6000	0.7452	**0.8630**	0.5652	0.6831	0.2004	0.2356
Linear SVM	0.5782	0.5578	0.7244	0.5571	0.6263	0.1566	0.1696
Ridge	0.5450	0.6182	0.5604	0.5432	0.5516	0.0901	0.0902
LDA	0.5450	0.6182	0.5604	0.5432	0.5516	0.0901	0.0902
Logistic Reg.	0.5440	0.6151	0.5576	0.5423	0.5498	0.0881	0.0882
Dummy	0.5005	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

Interestingly, when switching to TF-IDF features (Table 3), performance improved significantly across all models, with Ridge and LDA exceeding 0.87 in accuracy. This suggests that while LIWC captures psychological markers, the broader lexical variety captured by TF-IDF provides stronger discriminative signals for this task.

4.2. Model interpretability and feature analysis

Building on the selection of the Extra Trees Classifier for its balance of performance and transparency, we conducted a feature attribution analysis to identify the linguistic markers of depression. To better understand which linguistic features most influence predictions, we analyzed SHAP (SHapley Additive

Table 3: Classification results with TFIDF features on the CC2.

Model	Acc.	AUC	Rec.	Prec.	F1	κ	MCC
Ridge	**0.8783**	0.0000	0.8877	0.8713	0.8794	0.7566	0.7567
LDA	**0.8783**	0.9433	0.8877	0.8714	**0.8795**	**0.7566**	**0.7568**
LightGBM	0.8741	**0.9437**	0.8791	0.8704	0.8747	0.7482	0.7483
Extra Trees	0.8721	0.9430	0.8239	**0.9118**	0.8656	0.7442	0.7477
Random Forest	0.8430	0.9232	0.7890	0.8847	0.8341	0.6861	0.6902
Linear SVM	0.8371	0.0000	0.8283	0.8431	0.8356	0.6741	0.6743
Logistic Reg.	0.8354	0.8956	0.8461	0.8284	0.8371	0.6708	0.6710
AdaBoost	0.8021	0.8785	0.8553	0.7731	0.8121	0.6042	0.6077
Grad. Boosting	0.7979	0.8807	0.8347	0.7775	0.8051	0.5958	0.5975
Naive Bayes	0.7860	0.7877	**0.9140**	0.7277	0.8103	0.5720	0.5917
Decision Tree	0.7848	0.7767	0.7773	0.7891	0.7832	0.5696	0.5697
KNN	0.6048	0.6535	0.8916	0.5666	0.6928	0.2096	0.2559
QDA	0.5739	0.5739	0.1542	0.9597	0.2657	0.1477	0.2717
Dummy	0.5000	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

exPlanations) values and present a beeswarm plot in Figure 1.

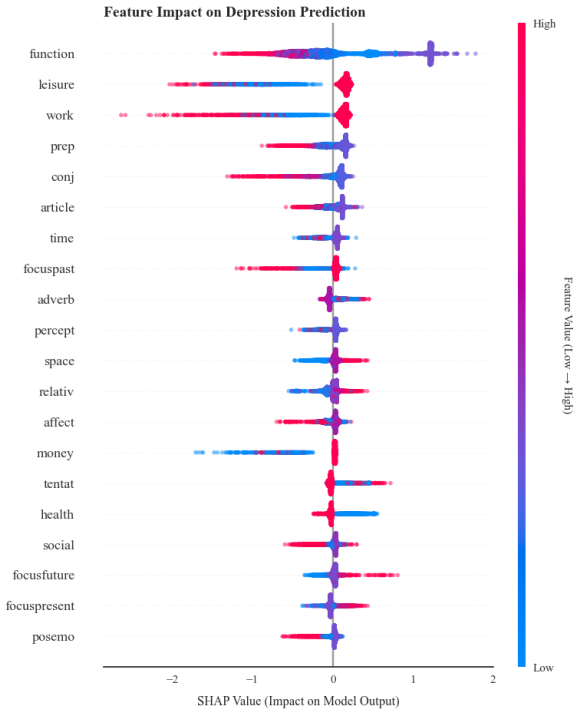

Figure 1: SHAP summary plot showing the top LIWC 20 features contributing to the prediction of the depressed class.

The beeswarm graph can be interpreted as follows. The visualization shows the SHAP values for every feature for every instance in each dataset, where each dot represents a SHAP value for a specific feature. The x-axis indicates the SHAP value magnitude, and the y-axis lists the features. In addition, the color of the dot reveals the value of the feature for that instance (e.g., red for high, blue for low). The features are ordered by the mean SHAP values. They are presented in a descending order of significance along the y-axis. A dot positioned on the right-hand side of the graph signifies a positive impact, while one on the left side indicates a negative impact. For instance, the LIWC-category 'work' showcases notably high negative SHAP values. The position of a data point along the x-axis signifies the degree of influence it exerts. Therefore, the further away from the point of origin (0), the greater the impact it has on the model's predictions.

The analysis reveals several critical insights:

- **Reduced engagement:** The LIWC categories `work`, `leisure`, and `social` exhibit high negative SHAP values. This indicates that a high frequency of words related to these domains is strongly associated with the non-depressed class. Clinically, this aligns with the social withdrawal and reduced interest in activities typical of depressive episodes.

- **Loss of drive and anhedonia:** The LIWC category `drives`, which include aspirations, achievements, and rewards, do not appear as prominent markers for the depressed class. This mirrors the clinical symptom of **anhedonia**, where individuals experience lowered expectations of reward and impaired reinforcement learning (Barch et al., 2016).

- **Temporal focus:** The `past tense` feature exerts a negative influence on the model's identification of the depressed class. This suggests that depressed individuals may focus less on past experiences and more on their current emotional state.

Considering that depression can lead to a lack of motivation, reduced interest in activities, and a sense of hopelessness, it is normal to expect that individuals face difficulties in pursuing their ambitions (Watson et al., 2020). Furthermore, the 'drives' LIWC category in LIWC 2015 version, includes also the 'achievements' category, which refers to experiences of success, and also the 'reward' category, which relates to the experience of receiving positive reinforcement. Many studies support that a core symptom of depression, usually referred as "anhedonia" is associated with lowered expectations of rewards and impaired reinforcement learning (Barch et al., 2016; Treadway and Zald, 2011).

Evaluation on the CC1 A second evaluation setting involved a comparison between the selected subset of the depression corpus and the randomly sampled control corpus. The subset from the depression corpus corresponded to a single user who served as a gold reference case. In particular, this user corresponded to an individual with a confirmed severe outcome, which was treated as a benchmark instance to approximate a real-world scenario of extreme depression risk. This setting should be interpreted as a stress-test scenario rather than a

general population model. By contrasting a confirmed severe case against a highly similar control user, we aim to evaluate whether linguistic signals remain detectable even under strong lexical similarity constraints.

To enable a fair comparison, we introduced a similarity-based control selection procedure. Cosine similarity was computed between vector representations of individual users and the gold reference user. Word embeddings were generated using the Greek FastText model [1]. The most similar non-depressed user achieved a similarity score of 0.99 but contained only 58 tweets, which was insufficient for reliable comparison. Therefore, the second most similar user (similarity = 0.97), with 12,076 tweets, was selected.

Tweets were preprocessed by removing URLs, usernames, punctuation, and emojis (via the demoji package), and by retaining only tweets containing more than two words. A random subsample of 10,000 tweets was selected to match previous experimental settings.

LIWC-based features Results using LIWC features (Table 4) show low classification performance, with accuracy ranging between 50% and 58%, only marginally above chance level. Most models achieved and MCC values close to zero, further indicating weak discriminative power. This behavior is attributed to the sparse coverage of the LIWC lexicon in the Greek Twitter corpus, resulting in sparse feature representations. Consequently, LIWC-based results were not further analyzed.

TF-IDF-based features TF-IDF features were then employed (Table 5). In contrast to LIWC, TF-IDF representations substantially improved performance across models. The Extra Trees classifier achieved the best overall results (Accuracy = 0.7150, F1 = 0.7332, = 0.4300, MCC = 0.4340), corresponding to an improvement of approximately 10 percentage points in accuracy compared to LIWC-based models.

Although TF-IDF features significantly outperform LIWC, several models exhibit noticeable discrepancies between Precision and Recall, suggesting instability in class-wise behavior. In particular, some classifiers favor recall at the expense of precision, indicating potential class bias. Further investigation is required to better understand the factors contributing to this imbalance.

4.3. Transformer-based models

We evaluated the performance of four pretrained models: (i) `bert-base-greek-uncased-v1`, (ii) `greeksocialbert-base-v2`, (iii) `bert-base-multilingual-cased` (mBERT), and (iv) `xlm-roberta-base`. This evaluation focuses on the CC2 dataset, which utilizes a control group constructed via topic-oriented extraction to ensure thematic consistency. The comparative metrics are summarized in Table 4.3.

[1] https://huggingface.co/facebook/fasttext-el-vectors

Table 4: Model performance with LIWC features; gold depressed vs random control user.

Model	Acc.	AUC	Rec.	Prec.	F1	κ	MCC
KNN	**0.5865**	0.6103	0.5858	0.5867	0.5862	**0.1730**	**0.1730**
Logistic Reg.	0.5801	**0.6187**	0.5977	0.5795	0.5851	0.1602	0.1625
MLP	0.5788	0.6178	0.6477	0.5719	**0.6043**	0.1575	0.1610
Extra Trees	0.5511	0.5743	0.5872	0.5556	0.5616	0.1022	0.1065
Linear SVM	0.5432	0.0000	0.4612	0.5533	0.4957	0.0865	0.0894
LDA	0.5137	0.5236	**0.7192**	0.5099	0.5965	0.0275	0.0297
Decision Tree	0.5076	0.5168	0.3295	0.5898	0.3197	0.0153	0.0225
AdaBoost	0.5066	0.5127	0.6022	0.5298	0.4076	0.0133	0.0649
Random Forest	0.5037	0.5636	0.6222	0.4899	0.4401	0.0075	0.0463
Naive Bayes	0.5024	0.5992	0.0092	0.6242	0.0182	0.0047	0.0238
QDA	0.5022	0.6000	0.0082	0.6368	0.0162	0.0045	0.0243
Dummy	0.5000	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

Table 5: Model performance with TFIDF features; gold depressed vs randomly selected user.

Model	Acc.	AUC	Rec.	Prec.	F1	κ	MCC
Extra Trees	**0.7150**	0.7804	0.7831	0.6894	**0.7332**	**0.4300**	**0.4340**
MLP	0.7113	**0.7824**	0.7581	0.6933	0.7240	0.4225	0.4247
Random Forest	0.6981	0.7722	0.7969	0.6656	0.7252	0.3962	0.4044
Linear SVM	0.6880	0.0000	0.7744	0.6605	0.7125	0.3759	0.3823
Logistic Reg.	0.6850	0.7450	0.6909	0.6830	0.6868	0.3700	0.3702
Naive Bayes	0.6764	0.6838	0.5772	**0.7200**	0.6406	0.3528	0.3600
Decision Tree	0.6566	0.6612	0.7084	0.6416	0.6733	0.3131	0.3150
AdaBoost	0.6375	0.7077	**0.9216**	0.5877	0.7176	0.2750	0.3349
KNN	0.6175	0.6480	0.7181	0.5982	0.6516	0.2350	0.2416
LDA	0.5986	0.6124	0.5944	0.5991	0.5966	0.1972	0.1973
QDA	0.5456	0.5456	0.7856	0.5324	0.6285	0.0913	0.1108
Dummy	0.5000	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000

Model	Acc.	Prec.	Rec.	F1
Greek BERT	0.948	0.938	**0.960**	0.949
GreekSocialBERT	0.942	0.934	0.953	0.942
mBERT	**0.953**	**0.965**	0.940	**0.952**
XLM-Roberta Base	0.944	0.961	0.926	0.943

Table 6: Model performance on the CC2.

Evaluation on the CC2 All transformer models achieved strong performance, with accuracy ranging from 94.2% to 95.3%. The multilingual BERT model obtained the highest overall accuracy (0.953), while Greek BERT achieved slightly higher recall (0.960), indicating strong sensitivity to depressed instances.

For Greek BERT, recall exceeds precision by approximately 2.2 percentage points (0.960 vs 0.938). In the context of depression detection, prioritizing recall may be desirable, as minimizing false negatives (i.e., failing to identify depressed individuals) is often considered critical. In contrast, multilingual BERT achieves higher precision (0.965) but slightly lower recall (0.940), suggesting a more conservative classification strategy. This model produces fewer false positives but may miss a greater num-

ber of true depression cases compared to Greek BERT.

Overall, these transformer architectures substantially outperform the traditional feature-based baselines presented in Section 4.1. This performance leap demonstrates the effectiveness of contextualized embeddings in capturing the nuanced semantic and syntactic structures associated with mental health discourse.

Evaluation on the CC1 We next evaluated the transformer models under the second experimental setting, comparing the gold depressed user with the cosine-similarity-matched control user (Table 4.3).

Performance decreases notably in this constrained setting, with accuracy ranging from 0.554 (`xlm-roberta-base`) to 0.779 (`greeksocialbert-base-v2`). GreekSocialBERT achieved the best overall performance (F1 = 0.776), followed by mBERT (F1 = 0.758). While mBERT showed a slightly more balanced precision-recall trade-off, GreekSocialBERT's superior scores suggest that its pretraining on social media data may provide an advantage when distinguishing subtle linguistic nuances in informal Greek text.

Model	Acc.	Prec.	Rec.	F1
Greek BERT	0.753	0.766	0.736	0.748
GreekSocialBERT	**0.779**	**0.788**	**0.767**	**0.776**
mBERT	0.759	0.761	0.758	0.758
XLM-Roberta base	0.554	0.717	0.534	0.420

Table 7: Model Performance on the CC1.

The significant performance drop compared to the CC2 experiment is expected and directly attributable to the similarity-based sampling procedure. Because the control user was selected based on extreme cosine similarity (0.97), the lexical and distributional overlap between the two classes is substantial. This minimizes "topic-driven" separability, where a model might simply distinguish between "talking about sadness" vs. "talking about sports", and forces the model to rely on much more granular linguistic markers.

This experiment serves as a stress-test scenario, evaluating whether models can maintain discriminative power when surface-level similarity is high. The overlap in vector space likely causes the transformer models to struggle with defining clear decision boundaries, as the embeddings for both classes occupy nearly identical regions

We avoid attributing the performance drop solely to overfitting, as the primary cause appears to be reduced inter-class variance due to similarity constraints. Moreover, while transformer models achieve strong predictive performance, they offer limited interpretability compared to feature-based approaches. Because classification decisions are based on high-dimensional contextual representations, isolating specific linguistic markers that distinguish depressed from non-depressed users remains challenging.

Moreover, while transformer models offer superior predictive power, this experiment highlights a critical trade-off: performance vs. interpretability. Unlike the feature-based analysis (LIWC/SHAP), the decision-making process of these high-dimensional contextual models remains opaque. While their performance is better in comparison to traditional models, they cannot explicitly reveal which Greek linguistic cues were the deciding factors.

5. Conclusions

We evaluated a range of machine learning (ML) and deep learning (DL) architectures for the binary classification of depressed versus non-depressed individuals. Our results identify mBERT as the overall top-performing model, achieving an accuracy and F1-score of 0.95 on the topic-oriented corpus (CC2). However, performance decreases substantially in the stress-test evaluation (CC1), where models are required to distinguish a gold-standard depressed user from a linguistically similar control user. This drop is likely due to the increased difficulty of the task, as the control user was selected based on high cosine similarity (0.97) to the target user. This high degree of lexical and semantic similarity may reduce class separability and increase classification ambiguity, suggesting that surface-level representations may be insufficient in highly controlled matching scenarios and that more fine-grained linguistic features may be beneficial.

For the ML baselines, we contrasted TF-IDF and LIWC features. TF-IDF achieved the highest predictive performance, with the Extra Trees (ET) classifier reaching an accuracy of 0.86 and an F1-score of 0.85. However, LIWC features offered greater interpretability, enabling direct mapping of linguistic patterns to meaningful LIWC categories.

Despite the more moderate performance of LIWC-based models (e.g., ET achieving 0.65), we utilized the ET classifier for our primary feature analysis due to its inherent robustness and the interpretability of its decision paths. LIWC features as a part of the feature engineering component has been exploited in numerous studies (Coppersmith et al., 2015; Resnik et al., 2013; Asgari et al., 2017; Tadesse et al., 2019). Their incorporation alongside other features has consistently demonstrated their potential to enhance classification performance, underscoring their significant role in capturing language-specific cues related to

depression. In this study, the interpretability analysis identified several prominent linguistic markers within the Greek depression corpus: (i) social withdrawal, manifested as a marked decrease in discourse related to social and leisure activities; (ii) professional disengagement, characterized by a significant reduction in work-related engagement; (iii) diminished drive, reflected in lower motivational cues and fewer references to achievements or rewards; and (iv) altered temporal orientation, evidenced by shifting patterns in time-reference words relative to the control group. These findings align with established clinical literature and support the existence of language-specific cues in depressive communication.

Ultimately, this study advocates for grounding automated screening in clinically validated evidence. While digital self-disclosure offers a useful proxy for risk, verifying these patterns against formal diagnoses remains a priority for future work. Establishing a clinically validated “ground truth” is essential to ensure that detected pattern, such as anhedonia, social withdrawal, and reduced engagement, accurately reflect the underlying depressive condition.

6. Copyrights

7. Acknowledgements

This work is part of the first author’s doctoral thesis. «The implementation of the doctoral thesis was cofinanced by Greece and the European Union (European Social Fund-ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning» in the context of the Act “Enhancing Human Resources Research Potential by undertaking a Doctoral Research” Subaction 2: IKY Scholarship Programme for PhD candidates in the Greek Universities».

8. Ethical considerations

The use of social media data for mental health screening raises significant ethical challenges regarding privacy, data ownership, and potential harm. Our study adheres to the following ethical principles: (i) data privacy and anonymization: Although the data used in this study (CC1 and CC2) were collected from public social media platforms, we recognize that users may not have intended for their posts to be used in a psychiatric research context. To protect user identity, all personal identifiers (i.e., usernames, locations etc.) were removed, (ii) this research is strictly observational, we did not engage with any users during the data collection process, and (iii) the models are intended as screening aids for researchers and clinicians, not as definitive diagnostic tools.

Beyond privacy, additional ethical concerns arise from the potential deployment of such systems. In particular, diagnostic errors, such as false positives and negatives, may lead to unintended consequences, including unnecessary stigmatization of users or failure to identify individuals who may require support.

9. Limitations and Future Work

Despite the contributions of this work, several limitations should be considered when interpreting the results:

1. Platform bias and corpus representativeness: our dataset is derived exclusively from social media (X/Twitter). These platforms tend to skew toward specific demographics, often younger individuals, which may not fully represent the linguistic patterns of the broader speaking population or those with different socio-economic backgrounds.

2. "Silver standard": we acknowledge that labels are based on self-reported diagnoses. While these provide a "silver standard" for training, they lack the clinical precision of a formal psychiatric evaluation conducted by a mental health professional.

3. Linguistic resource constraints: While we utilize LIWC to bridge the explainability gap, it is important to note that the Greek version of the LIWC dictionary, while robust, may not capture the full nuanced range of informal "internet slang" in the Greek digital landscape as effectively as the original English version.

4. Stress-test generalizability: The stress-test experiment, which includes a single user in the

depression class, may capture idiolectal (user-specific) linguistic patterns rather than generalizable markers of depression. As such, its results should be interpreted as exploratory.

Future work will extend explainability to transformer-based models using methods such as SHAP or integrated gradients, to better understand which features contribute to depression prediction.

10. Bibliographical References

Moez Ali. 2020. *PyCaret: An open source, low-code machine learning library in Python*. PyCaret version 1.0.0.

Salma Almouzini, Maher khemakhem, and Asem Alageel. 2019. Detecting arabic depressed users from twitter data. *Procedia Comput. Sci.*, 163(C):257–265.

Meysam Asgari, Jeffrey Kaye, and Hiroko Dodge. 2017. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228.

Deanna M. Barch, David Pagliaccio, and Katherine R. Luking. 2016. Mechanisms underlying motivational deficits in psychopathology: similarities and differences in depression and schizophrenia. *Current Topics in Behavioral Neurosciences*, 27:411–449.

Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2025. A survey on multilingual mental disorders detection from social media data. *CoRR*, abs/2505.15556.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Medicine*, 3.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of ICWSM*, pages 128–137.

Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preotiuc-Pietro, David A. Asch, and H. Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115:11203–11208.

Lucía Gómez-Zaragozá, Javier Marín-Morales, Mariano Alcañiz, and Mohammad Soleymani. 2025. Speech and text foundation models for depression detection: Cross-task and cross-language evaluation. *Interspeech 2025*.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online. Association for Computational Linguistics.

Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen. 2018. Deep learning for depression detection of Twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 88–97, New Orleans, LA. Association for Computational Linguistics.

Glorianna Jagfeld, Fiona Lobban, Paul Rayson, and Steven Jones. 2021. Understanding who uses Reddit: Profiling individuals with a self-reported bipolar disorder diagnosis. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 1–14, Online. Association for Computational Linguistics.

Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring tweets for depression to detect at-risk users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 32–40, Vancouver, BC. Association for Computational Linguistics.

Danxia Liu, Xing Lin Feng, Farooq Ahmed, Muhammad Shahid, and Jing Guo. 2022. Detecting and measuring depression on social media using a machine learning approach: Systematic review. *JMIR Ment Health*, 9(3):e27244.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates, Austin, TX.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A

python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353, Seattle, Washington, USA. Association for Computational Linguistics.

Wesley Ramos dos Santos, Rafael Lage de Oliveira, and Ivandré Paraboni. 2023. Setembrobr: a social media corpus for depression and anxiety disorder prediction. *Lang. Resour. Eval.*, 58(1):273–300.

Vivian Stamou, George Mikros, George Markopoulos, and Spyridoula Varlokosta. 2024. Establishing control corpora for depression detection in Modern Greek: Methodological insights. In *Proceedings of the Fifth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, pages 68–76, Torino, Italia. ELRA and ICCL.

Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Michael T. Treadway and David H. Zald. 2011. Reconsidering anhedonia in depression: lessons from translational neuroscience. *Neuroscience and biobehavioral reviews*, 35 3:537–55.

Rebecca Watson, Kate Harvey, Ciara McCabe, and Shirley Reynolds. 2020. Understanding anhedonia: A qualitative study exploring loss of interest and pleasure in adolescent depression. *European Child & Adolescent Psychiatry*, 29(4):489–499.

Sourabh Zanwar, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2023. SMHD-GER: A large-scale benchmark dataset for automatic mental health detection from social media in German. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1526–1541, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhenwen Zhang, Jianghong Zhu, Zhihua Guo, Yu Zhang, Zepeng Li, and Bin Hu. 2024. Natural language processing for depression prediction on sina weibo: Method study and analysis. *JMIR Mental Health*, 11.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Zhu. 2022. Symptom identification for interpretable detection of multiple mental disorders on social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9970–9985, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.