Faculty of Mathematics
and Computer Science

Tehnici de realizare a sistemelor inteligente (TRSI 2024)

Cross-Linguistic Analysis for Depression Detection

Dinică Mircea

Department of Computer Science, Babeş-Bolyai University
1, M. Kogalniceanu Street, 400084, Cluj-Napoca, Romania
E-mail: .....

Abstract

The purpose of this paper is to research the viability of machine translation of text from English to Romanian, in order to use Natural Language Processing(NLP) and Artificial Intelligence(AI) to detect if the text shows signs of depression or not.

This paper describes its content in six chapters. Chapter 1 provides the motivation and objective for the depression detection application. Chapter 2 gives details regarding the dataset used and the pre-processing techniques, namely Linguistic Word Inquiry and Word Count which recognizes the emotions and grammar parts which the tokens belong to. It also describes why the machine learning classification algorithm Random Forest was chosen and the evaluation metrics used to analyse the performance. Chapter 3 details the results of the training for two experiments on the original dataset in English. Chapter 4 describes the results achieved when translating the dataset in Romanian and training the machine learning classifier using the same methodology as in . Chapter 5 details a SWOT analysis and the final chapter outlines the conclusions and proposes future development ideas.

1. Introduction

1.1. Motivation

Depression stands as a mental health affliction with profound impacts on both psychological and physical well-being. Characterized by a disinterest in routine activities, sleep disturbances, inability to feel pleasure, and in severe cases, thoughts of suicide [3], it has become a problem across worldwide. Furthermore, individuals with major depressive disorder face a bigger risk of cardiovascular problems, not optimal treatment outcomes, and higher rates of morbidity and mortality [6].

The World Health Organization (WHO) identifies depression as the primary contributor to global disability, affecting over 300 million individuals worldwide [10]. Particularly alarming is the revelation that adolescents with severe depression are 30 times more prone to suicide [11]. Although depression is a big problem worldwide, it is not known

exactly what causes it. Cultural, psychological, and biological factors play a part, but how they all fit together was not discovered.[4].

1.2. Objective

The purpose is to achieve two things. Firstly, provide an accessible application of identifying individuals who may be experiencing symptoms of depression. By analyzing the language used in written communications, such as social media posts, emails, or chat messages, the tool aims to offer an initial assessment of an individual's mental well-being. This approach can facilitate early intervention and support, potentially preventing more severe depressive symptoms and their associated consequences.

Secondly, evaluate the performance of the machine learning model in a cross-linguistic context. To achieve this, the dataset will be translated dataset from English to Romanian and assess the performance of the model on both language versions. This comparative analysis will tell if of the tool is accurate across different languages and cultural contexts.

With this study the hope is to contribute to the advancement of computational techniques for mental health assessment and intervention. The aim is to provide clinicians, researchers, and individuals themselves with a valuable resource for early detection and prevention of depression, ultimately encouraging improved mental well-being and quality of life.

2. Model Analysis from Data Preprocessing to Evaluation

In order to develop an accurate AI system for detecting depression from textual data, every step must be analysed carefully. This chapter describes the dataset, pre-processing techniques, the process of selecting the model for this task and the evaluation metrics used.

2.1. Dataset Overview

The investigation relies on a carefully compiled dataset [8], crafted in order to advance in mental health classification research. Gathered through web scraping techniques from diverse Subreddits, this dataset contains discussions and viewpoints on mental health topics. The aim of creating this dataset was to examine textual patterns which indicate depression's presence or absence in individuals, as seen from their online conversations .

The raw data was sourced by employing web scraping techniques, targeting specific Subreddits known for their discussions on mental health issues. This approach ensured that the data collected was relevant to the research objectives, capturing a diverse range of experiences and expressions related to mental health.

Comprising 7,650 unique entries, the dataset is enough for an accurate machine learning algorithm. Each entry is annotated with an is_depression label, distinguishing between texts that indicate the presence of depression (labeled '1') and those that do not (labeled '0'). This labeling process was carried out with careful consideration to ensure accuracy and reliability in the classification [8].

A noteworthy aspect of the dataset is its well-balanced nature, with 3,900 entries labeled as non-depression and 3,831 entries indicating depression. This balance is important in avoiding bias in the predictive modeling process, ensuring that the resulting classification model is accurate.

2.2. Pre-Processing Tool

In natural language processing, especially within the context of psychological research, the tool chosen to process and interpret the data is as important as the data itself. For this reason, exploration of the dataset uses the latest version of a text analysis software, LIWC-22 (Linguistic Inquiry and Word Count).

The tool is able to analyze language systematically, overcoming the complexities that early computer-based text analysis methods encountered [1].With LIWC-22, researchers have at their disposal a software tool that not only takes from previous versions but also incorporates the latest advances in text analysis. Its expanded dictionary and enhanced software capabilities make it possible to analyze language samples with depth and precision. This core component, comprising over 12,000 words, word stems, phrases, and select emoticons, is organized into categories

and subcategories designed to capture a wide array of feelings. This arrangement allows for a accurate analysis of text, offering insights into the psychological state, social relationships, and cognitive processes of individuals based on their word usage.

The LIWC-22 Dictionary has a hierarchical organization, where words are not only categorized but also interlinked across multiple dimensions. For instance, the word ”cried” contributes to categories such as emotion, sadness, and past focus, illustrating the dictionary's complexity and depth. This structure enables LIWC-22 to provide a comprehensive analysis of text, reflecting various emotional and cognitive dimensions [1].

In the table 1 there are examples of the words linked to their coressponding categories and it can be seen that the words represent specifics of their category.

Social	**Culture**	**Lifestyle**	**Physical**
admiration	norwegian	free time	abs
company	nuclear	accomplish	aerobic
listener	online	real estate	ailment
locals	arabic	gaming	alcohol
refugee	political	qualify	deaf
reassure	phonecall	amusement	death
trust	person of color	god	kidney
tweets	racist	remodel	lactose
twins	bill of rights	art	salad
uncle	scanner	greed	ketogen
loyal	bots	rent	depressed
commitment	candidate	assignment	diabet
confess	opposition party	psychologist	sauna

Table 1. Examples of categories and its words in LIWC-22

2.3. Reasoning Behind Choosing Random Forest

In machine learning, selecting the most appropriate algorithm is very important to the success of any classifying task. This is also true in the case of depression detection, where the complexity and variability of the data demand an approach that is not only accurate but also can work on texts from different cultures. Looking at a comprehensive study that evaluated twelve distinct machine learning algorithms across seven datasets[9], it was decided to use Random Forest (RF) as the AI model.

The study [9] in question compared the performance of several algorithms, including Naive Bayes (NB), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Artificial Neural Networks (ANN), Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), Hoeffding Tree (HT), Decision Tree (DT), C4.5, Classification and Regression Tree (CART), Random Forest (RF), and Bayesian Belief Networks (BB), across multiple metrics. Among these, Random Forest showed the most consistent and high results, showing superior accuracy, precision, and Matthew's Correlation Coefficient (MCC). Following Random Forest, the algorithms of Neural Networks (NN), Naive Bayes (NB), Bayesian Belief Networks (BB), and Logistic Regression (LR) were identified as the next most effective, in descending order of accuracy.

The study [9] also highlighted the significance of the kappa statistic and Root Mean Square Error (RMSE) as important factors in assessing model performance, further validating the consistency of Random Forest in handling diverse and complex datasets. With these statistics, and in accordance with the study's conclusion, the selection of Random Forest is motivated by its results across multiple validation metrics.

In the context of this study focused on depression detection, the dataset resembles the Breast Cancer Wisconsin dataset, because the model is also a binary classifier. However, the model differentiates itself with a higher dimensionality, processing 119 input attributes, which poses a greater complexity in feature representation and selection. For this dataset (Random Forest) RF achieved the highest accuracy at 97.85%, suggesting it was the most successful in correctly identifying cases of breast cancer. It also had the highest kappa value of 95.03%. Precision with RF was

great as well, hitting a high of 98%, while its recall was nearly as impressive at 97.9%, showing its ability to identify most of the positive cases.

2.4. Evaluation Metrics

Metrics are a crucial part of evaluating the effectiveness of a binary classifier. It's important to use a variety of tools and methods to understand different aspects of the model's performance. Here the metrics that were chosen for the evaluation of the depression binary classifier:

- **Classification Metrics**: These include accuracy, precision, recall, and the F1-score, which together provide a comprehensive overview of overall model performance. Their respective equations are detailed, where TP are true positives, TN are true negatives, FP are false posivites and FN are false False Negatives.
  - **Accuracy** measures the overall correctness of the model across all predictions 1.
  - **Precision** assesses how many of the positively predicted cases were actually positive 2.
  - **Recall** (or sensitivity) determines how many of the actual positive cases were correctly identified by the model 3.
  - **F1-Score** is the harmonic mean of precision and recall, helping balance the two in scenarios where one may be more important than the other 4.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

- **Confusion Matrix**: This is a table that visualizes the performance of the binary classifier by showing the actual versus predicted classifications. It helps identify the kinds of errors the model is making, such as confusing one class for another.
- **ROC Curve**: This graph shows the ability of the model to distinguish between the two classes at various threshold levels. It plots the true positive rate against the false positive rate, providing insight into the trade-offs between capturing positives and avoiding false alarms [5].
- **Feature Importance**: This metric highlights which inputs or variables in your data have the most influence on the model's predictions. Understanding feature importance can help in refining the model by focusing on the most relevant factors.

By using these metrics, a detailed understanding of your model's strengths and weaknesses can be achieved, guiding improvements and ensuring it performs well across various conditions.

3. Model Selection and Hyperparameter Tuning

3.1. Model Training Approach

To achieve an optimal model, it is essential to explore various training methodologies. This chapter describes the procedures followed in training the model.

3.2. Initial Model Training Strategy

In the initial training phase, all available features were utilized without any adjustments to hyperparameters. The default settings of the Random Forest Classifier from sklearn were applied [2]. The dataset was partitioned in a 75/25 train/test split, ensuring an equal distribution of positive and negative cases by using the stratify option.

The first experiment's results reveal a strong performance across the chosen metrics. The Classification Metrics plot shows high values for Accuracy (0.96), Precision (0.99), Recall (0.93), and F1 Score (0.96), indicating an efficient model with a balanced approach to both relevance (precision) and completeness (recall) 1.

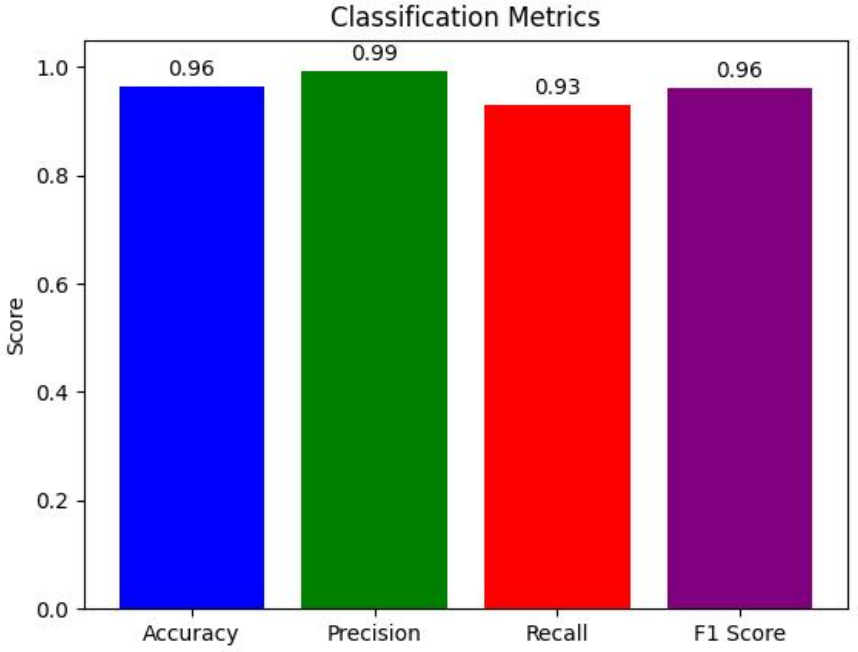

Fig. 1. Classification Metrics First Experiment

The Confusion Matrix provides a visual confirmation of the model's performance, with a high number of true positives (891) and true negatives (969), and relatively few false positives (6) and false negatives (67) 2. This suggests the model has more problems when finding depression.

Looking at the ROC Curve, the model demonstrates an excellent ability to distinguish between the classes, as evidenced by the area under the curve (AUC) being close to 1 (0.99) 3. This suggests that the model has a good discrimination capability with a high true positive rate and a low false positive rate across different thresholds.

The Top 10 Feature Importances plot indicates which features have the most influence on the model's predictions. The leading features, labeled as 'WC' (Word Count) and 'WPS' (Words per Sentence), seem to be the most significant drivers, with the others contributing to varying lesser degrees 4. This shows that the length of the given text is very important in order for the model to give an accurate prediction.

Overall, the model appears to be highly effective, with strong performance indicators, which seems to be the result of the analysis done for choosing the pre-processing methods and classifier, namely LIWC [1] and Random Forrest.

3.3. Hyperparameter Tuning

The selection of Random Forest hyperparameters is guided by the study [7]. The adjusted hyperparameters are:

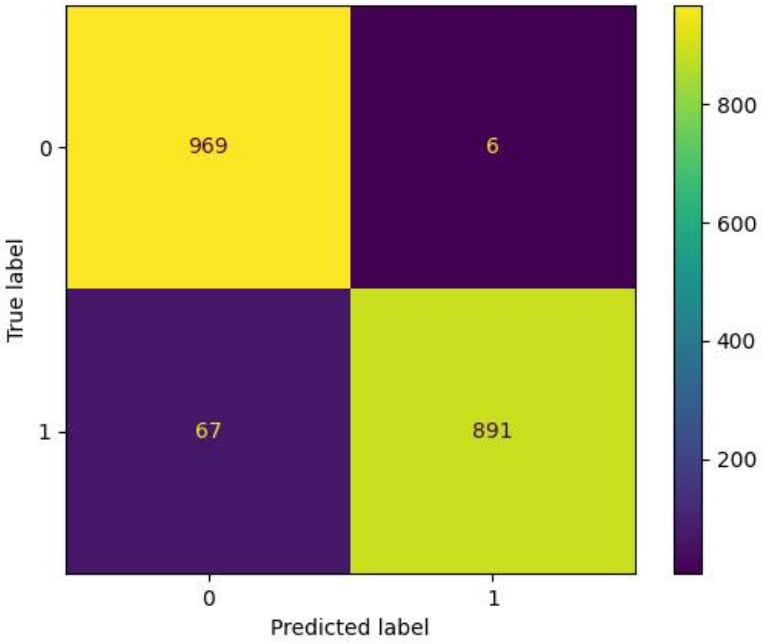

Fig. 2. Confusion Matrix First Experiment

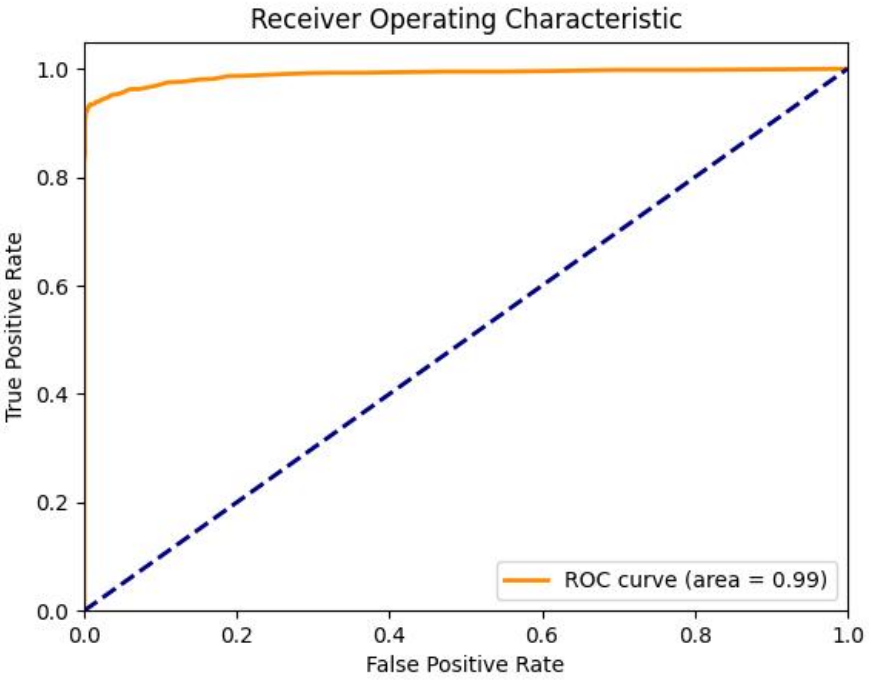

Fig. 3. ROC Curve First Experiment

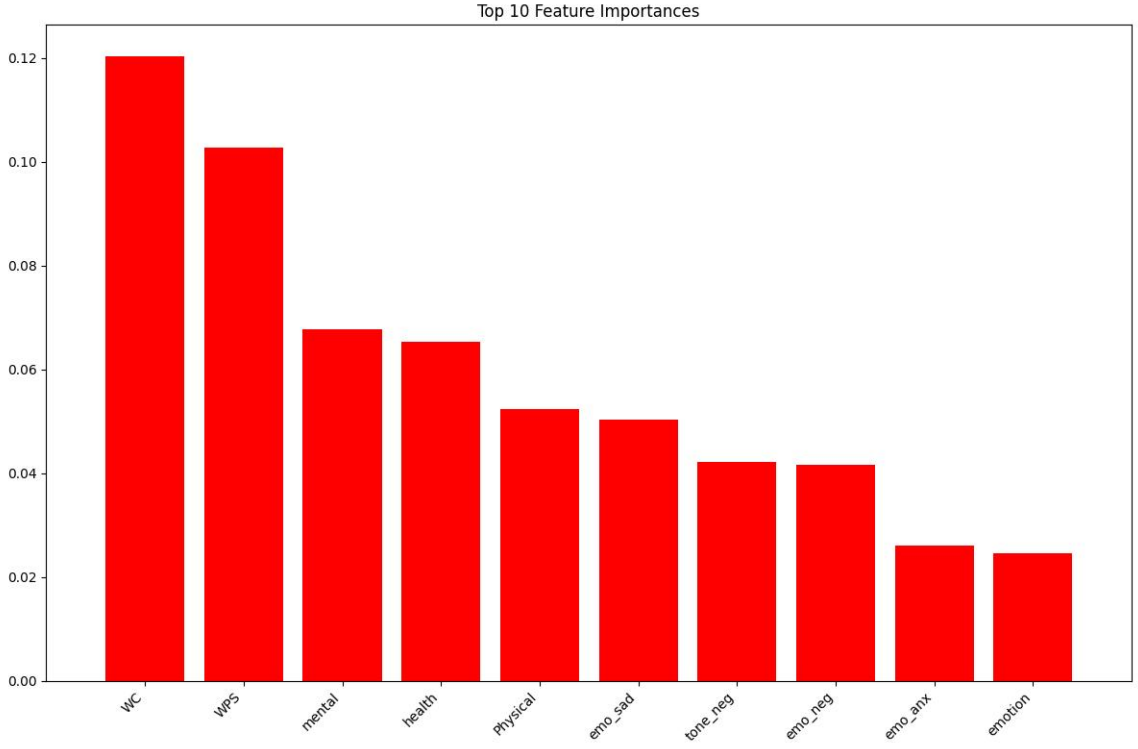

Fig. 4. Top 10 Feature Importances First Experiment

- **mtry**: This represents the number of features considered for splitting at each node. Lower values promote tree diversity and are beneficial when there are many relevant predictors. Given the large feature set, a value lower than the default square root of the number of features is suggested to prevent dominant features from overshadowing others.
- **Number of Trees**: A sufficient number of trees ensures stable predictions and importance estimates. While more trees generally improve model performance, beyond a certain point the marginal gains diminish. For practical purposes, between 500 to 1000 trees are recommended.
- **Node Size**: The node size controls the depth of the tree. Smaller node sizes can potentially lead to over-fitting, particularly when the number of features is high. A larger than 1 node size is preferred to mitigate this risk and improve computational efficiency.
- **Sample Size**: The proportion of data used for training each tree. Smaller sample sizes lead to more diversity but can decrease individual tree accuracy. Optimal sample size needs to be problem-specific but sampling a subset, such as between 20% and 90% of the data, can yield good results while reducing runtime.

These hyperparameters were tuned using Sequential Model-Based Optimization (SMBO) to determine their optimal values while considering the Area Under the ROC Curve (AUC) as the performance metric [7].

For the depression binary classifier multiple configurations were tested. It was noticed that the optimal ranges for the hyperparameters were:

- **mtry**: between 6 and 10
- **Number of trees**: between 700 and 1000
- **Node Size**: between 3 and 9
- **Sample size**: between 6 and 9

After training with all combinations of the mentioned ranges, the one who got the best results was 6 for mtry, 900 for number of trees, 3 for node size and 8 for sample size. The same metrics as in the first experiment were used to analyze the performance of the model and improvements were seen. For the classification metrics accuracy(0.97) has improved by 0.01, precision stayed the same, recall(0.96) was the one who improved the most by 0.03, and F1-score(0.97) improved by 0.01. These values can be seen comparing Figure 1, which shows the classification metrics for the first experiment, with Figure 5, which shows them for the second experiment. This shows that because of hyperparameters tuning, it succeeded in improving the part where the model from the first experiment lacked.

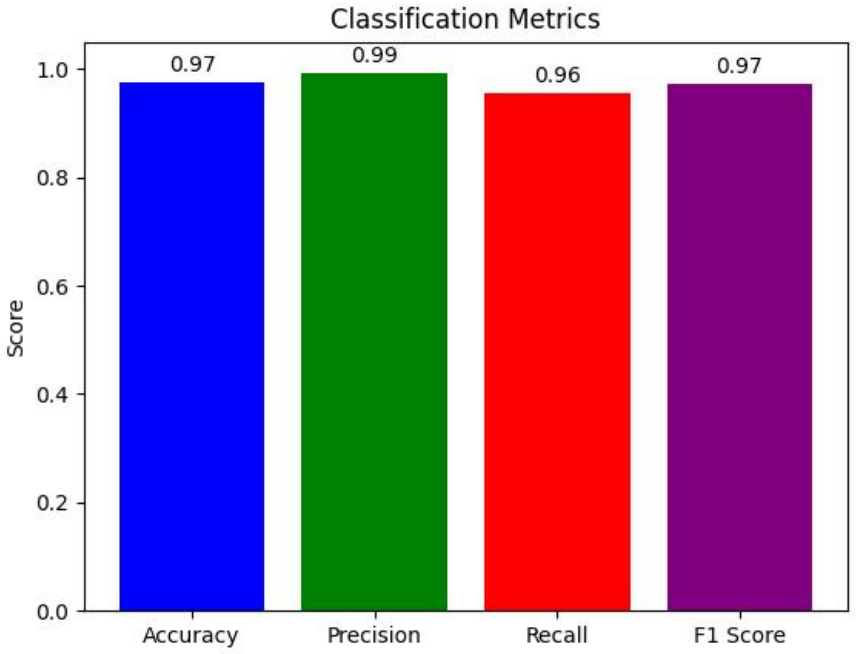

Fig. 5. Classification Metrics Second Experiment

In the case of the confusion matrix for the second experiment 6 it can be discovered that in comparison with the one for the first experiment 2, as can the improvement of recall also tell, the values of the false positives decreased, from 67 to 43.

For the ROC curve for this experiment 7, the area stayed the same when looking to the first two digits, but an improvement of the true positive rate can be noticed from the initial experiment 3.

Seeing the top 10 features of the second experiment 8, it was noticed that in comparison with the initial one 4 that even though WC (Word Count) and WS (Words per Sentence) are the still the most important features, they are now by much less, from 0.12 and 0.10 to both being at 0.09. This shows that now there are more features taken into account when the model makes the classification and each is more influential. Also it is remarked that a feature in the first ten was changed, namely emo_anx (anxiety) with cause (causation).

4. AI Model for Romanian Language

For the Romanian language model, the methodology was replicated consistently. The dataset was translated using ”googletrans” Python library [12] and LIWC served as the tool for preprocessing the data. However, the most recent English dictionary for LIWC-22 has not been translated into Romanian. The latest available version for Romanian is the LIWC-2015 dictionary, which contains only 86 features, compared to the 119 features available in the English version.

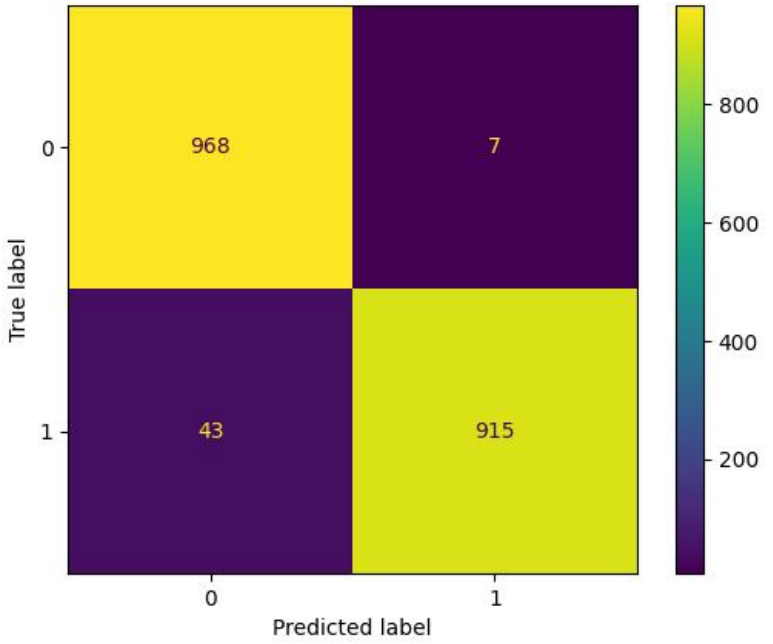

Fig. 6. Confusion Matrix Second Experiment

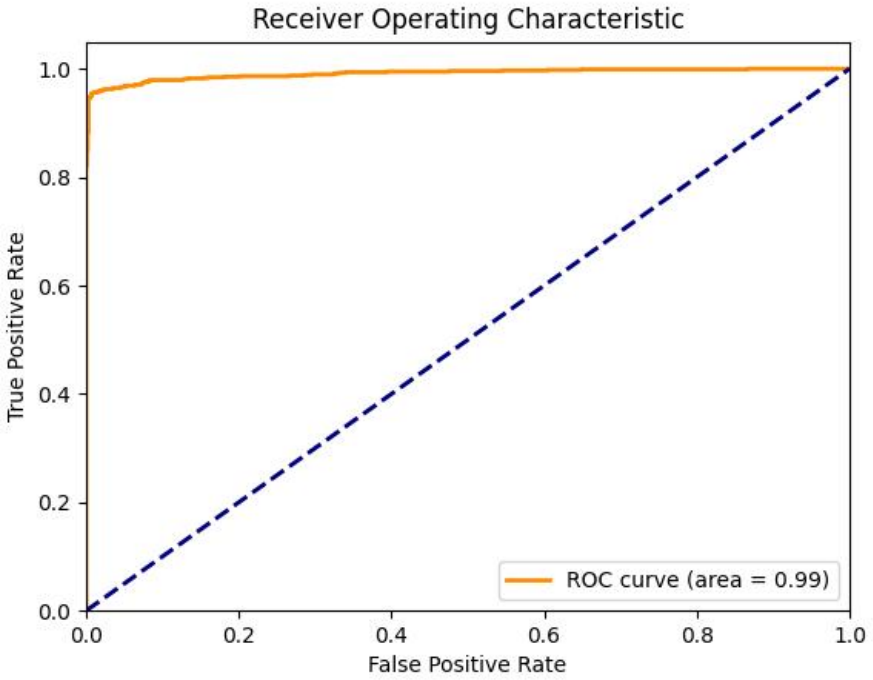

Fig. 7. ROC Curve Second Experiment

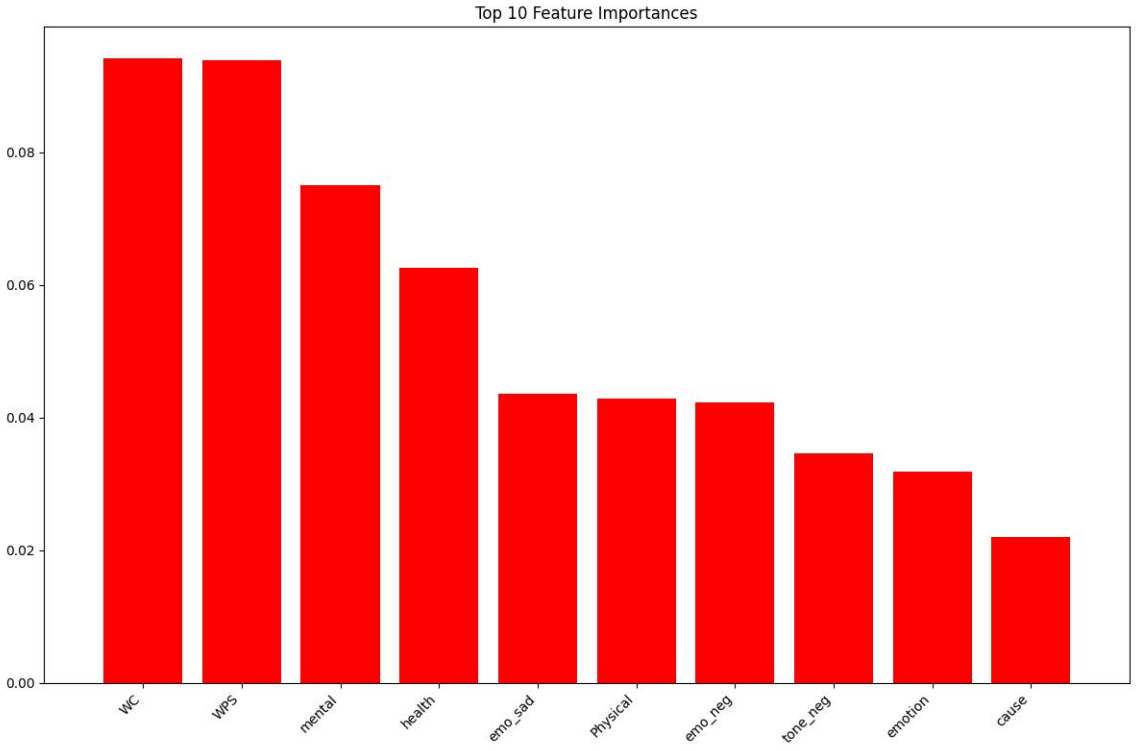

Fig. 8. Top 10 Feature Importances Second Experiment

The training approach was aligned with the methodology used for the English model during the second experiment. The hyperparameters and the proportion of training to testing data remain consistent:

- **train/test split**: Maintained at 75/25, using stratification based on the 'is_depression' label
- **mtry**: 6
- **Number of trees**: 900
- **Node Size**: 3
- **Sample size**: 8

This methodology facilitates a direct comparison between the performance of the Romanian and English models, ensuring consistent evaluation criteria across both. The same metrics were used for analysis. In terms of classification metrics, the most notable discrepancy arises in recall, where the Romanian model scores 0.87, falling short of the English model's 0.96 by 9 percentage points. Additionally, both accuracy and the F1-score have diminished by 0.05, while precision experienced the least impact, decreasing from 0.99 to 0.97. These metrics are illustrated in Figure 5 for the English model and in Figure 9 for the Romanian model.

The diminished recall in the Romanian model suggests it is less adept at identifying true positive cases as compared to the English model. This lower performance may come from the nuances lost during the translation of the dataset from English to Romanian using the Googletrans library [12]. Such translation challenges could contribute to the model's reduced effectiveness, highlighting the influence of linguistic or cultural differences on the model's ability to generalize across languages. The smaller reductions in accuracy and F1-score indicate that while the model is somewhat less effective overall, it still maintains a reasonable level of precision.

The confusion matrix 10 illustrates a significant increase in false positives, rising from 43 in the English model 6 to 123 in the Romanian model. However, the rise in false negatives was less marked, increasing from 7 to 26.

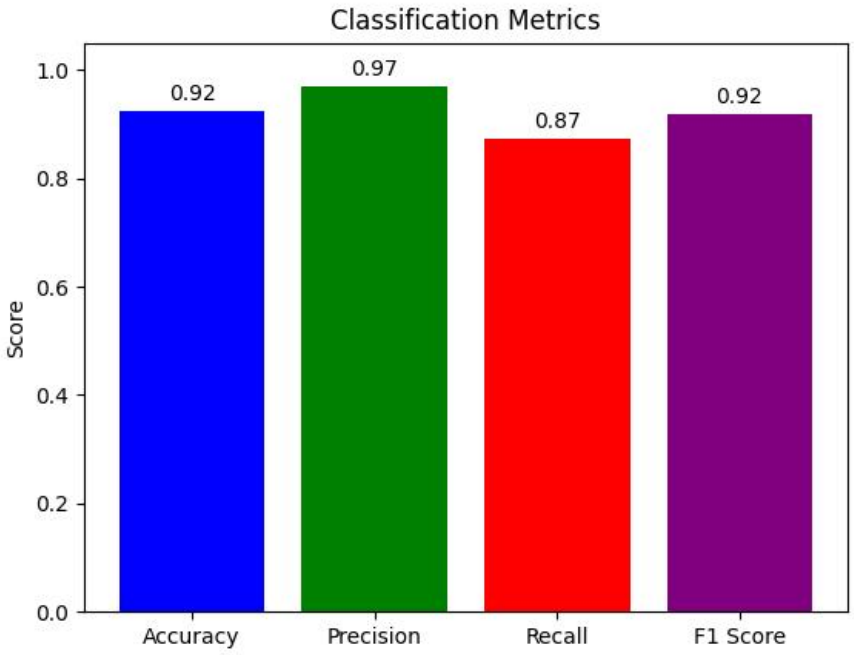

Fig. 9. Classification Metrics Romanian Model

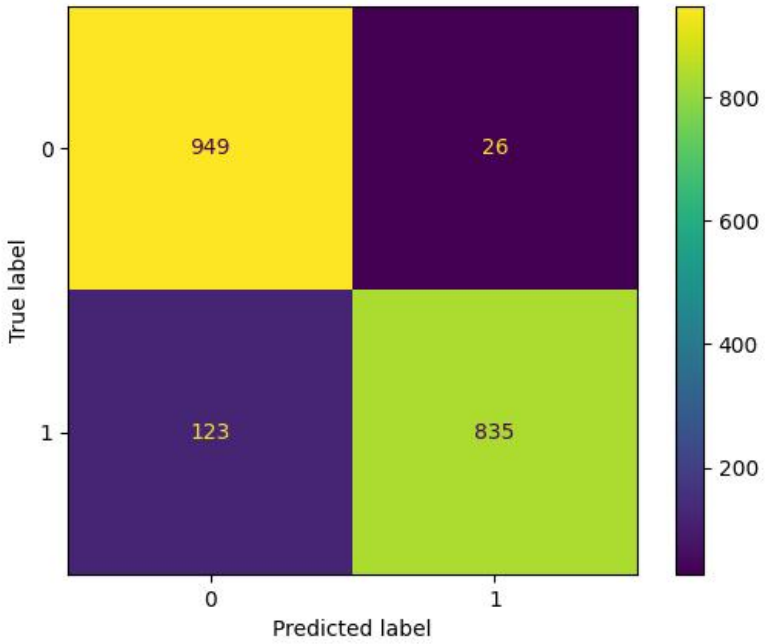

Fig. 10. Confusion Matrix Romanian Model

Regarding the ROC curve, the Area Under the Curve (AUC) experienced a slight decrease of 0.01, as depicted in Figure 11 compared to Figure 7. These metrics collectively indicate that the issues observed during the experimentation with the English model are more pronounced in the Romanian classifier.

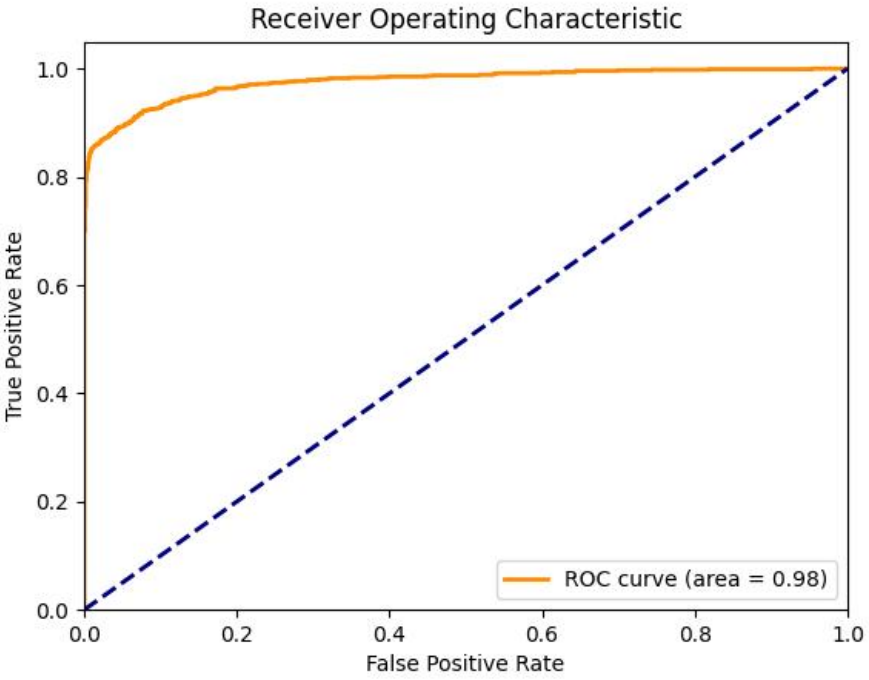

Fig. 11. ROC Curve Romanian Model

The most significant features for the Romanian classifier 12 align more closely with those observed in the initial experiment for the English model 4. Notably, Word Count (WC) has risen above 0.12, surpassing the prominence it held in the first experiment. This contrasts with the enhancements seen in the second English experiment 8, where a diminished reliance on WC and WPS (Words per Sentence) indicated a broader array of features influencing the English classifier's decisions. However, this diversification does not appear to extend to the Romanian model.

Some features remain consistent with the English LIWC-22 dictionary features listed in the top 10 for the second experiment; for instance, "sad" aligns with "emo_sad", and both "cause" and "health" are both present and "negemo" is the same as "emo_neg". The "anx" feature mirrors "emo_anx" from the first experiment's top features. The presence of "Period" suggests that the Googletrans library has introduced punctuation marks, which have become a significant element. The evolution from LIWC-2015 to LIWC-22 is further evidenced by the removal of the "interrog" category in the Romanian model's top 10 features, which was dropped in LIWC-22 due to its low base rates, internal reliability, or infrequent usage, as noted in [1]. The "ipron" feature might also result from the machine translation process.

5. Discussion

This chapter gives a SWOT analysis of the proposed model.

- **Strengths**:The English model showed better performance than the Romanian one, due to LIWC-22 being only in English, meaning that translating input text from any language to English may show reliable results.
- **Opportunities**: The model can be used in order to prevent severe symptoms of depression through analysing a company's web application text or even employees messages or posts. It could also greatly aid in targeted marketing for psychologists or enhancing mental health awareness.

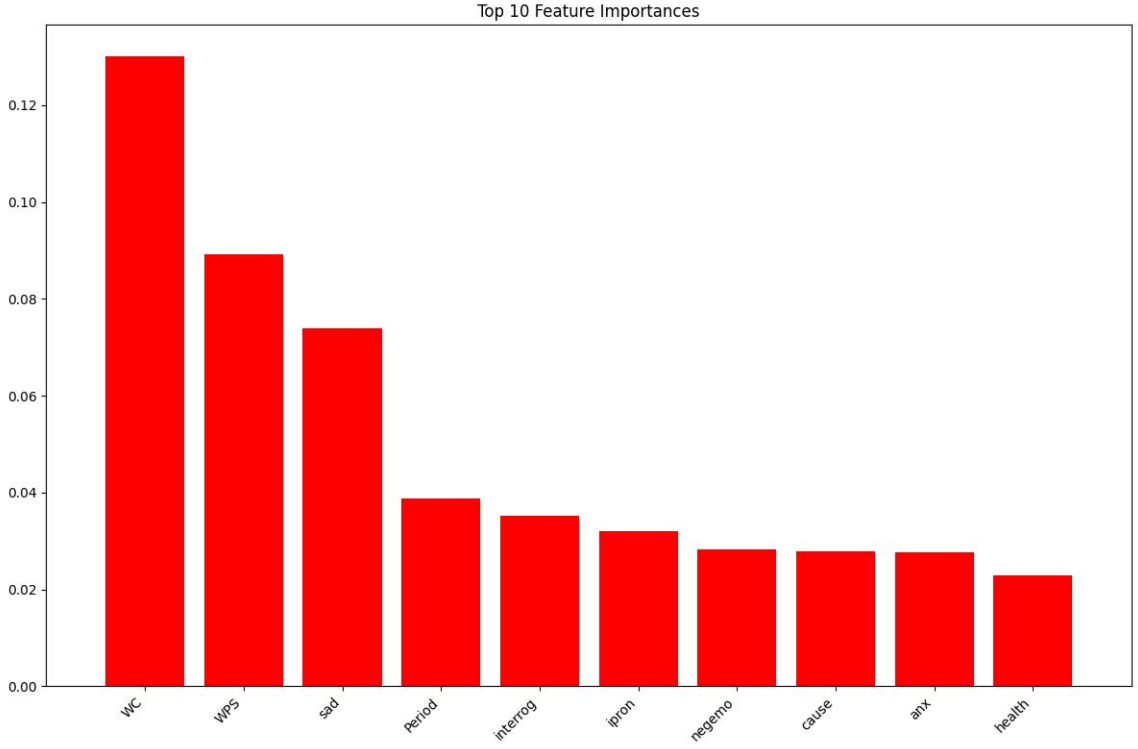

Fig. 12. Top 10 Features Romanian Model

- **Weaknesses**: LIWC [1] specifies its use strictly for academic purposes, any commercial application would need to adopt a different pre-processing approach. Also a professional psychologist is needed in order to give a human-based assesment.
- **Threats**: Due to the information being processed by the model being sensitive, security is very important in the process of parsing input from the an application's frontend.

6. Conclusions and future work

Developing a multilingual tool presents significant challenges. There is a much richer body of literature for English than for Romanian, which impacts the performance of AI models, as showed by the experiments. The English model achieved a precision of 96%, significantly higher than the Romanian model's 87%. This difference largely comes from the translation methods used and the limitations of the pre-processing tool LIWC. Despite using Google Translate, a leading translation service, the Googlelib [12] encountered difficulties in maintaining the original text's meaning. Additionally, the most recent version of LIWC, LIWC-22, is only available in English, which meant a downgrade to LIWC-2015 for Romanian, which is seven years behind in advancements, as reflected in the precision of the Romanian classifier.

For future improvements, employing an AI-based translation tool could preserve text meaning more effectively. The next step for the tool is its deployment to a production environment, transitioning from localhost to a cloud hosting platform to ensure broader accessibility and reliability.

Moreover, while communicating, words represent only a minor fraction of the information conveyed. The tool, which solely analyzes text, is insufficient to conclusively determine if a person is depressed. Therefore, it is important to note that the tool serves merely as a preliminary assessment of a person's emotional state, a thorough evaluation requires a professional in psychology. For a more accurate computer-based analysis, it would be necessary to consider all aspects of communication, both verbal and nonverbal.

7. Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author used ChatGPT in order to paraphrase. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

[1] Boyd, R.L., Ashokkumar, A., Seraj, S., Pennebaker, J.W., 2022. The development and psychometric properties of liwc-22. Austin, TX: University of Texas at Austin , 1–47.

[2] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G., 2013. API design for machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122.

[3] Cui, R., et al., 2015. A systematic review of depression. Curr Neuropharmacol 13, 480.

[4] Gross, M., 2014. Silver linings for patients with depression? Current Biology 24, R851–R854.

[5] Hoo, Z.H., Candlish, J., Teare, D., 2017. What is an roc curve?

[6] Luo, Y., Zhang, S., Zheng, R., Xu, L., Wu, J., 2018. Effects of depression on heart rate variability in elderly patients with stable coronary artery disease. Journal of Evidence-Based Medicine 11, 242–245.

[7] Probst, P., Wright, M.N., Boulesteix, A.L., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: data mining and knowledge discovery 9, e1301.

[8] Shinde, V., 2022. Depression: Reddit dataset (cleaned). https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned/data. Accesed: 12-03-2024.

[9] Siraj-Ud-Doulah, M., Islam, M.N., 2023. Performance evaluation of machine learning algorithm in various datasets .

[10] Smith, K., De Torres, I., 2014. A world of depression. Nature 515, 10–1038.

[11] Stringaris, A., 2017. What is depression?

[12] Suhun, H., 2020. Googletrans documentation, “googletrans: Free and unlimited google translate api for python — googletrans 3.0.0 documentation”,. https://py-googletrans.readthedocs.io/en/latest/. [Online; accesed 29-March-2024].