BABEŞ-BOLYAI UNIVERSITY CLUJ-NAPOCA
FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
SPECIALIZATION COMPUTER SCIENCE ROMANIAN

DIPLOMA THESIS

Depression signs detection

Supervisor
Lector Universitar, Lupea Mihaiela

Author
Dinica Mircea

2024

ABSTRACT

This scientific study delves into the realm of depression detection through the lens of artificial intelligence (AI) and natural language processing (NLP). With a focus on enhancing the accuracy and applicability of depression identification tools, the study first provides a ~~thorough~~ understanding of depression statistics and insights. It then outlines the primary objective: to develop a ~~robust~~ AI model capable of detecting depression while also assessing its performance across linguistic boundaries.

The chapters unfold to reveal a comprehensive exploration of the dataset, shedding light on its composition and characteristics. Subsequent chapters delve into the intricate process of model selection and hyperparameter tuning, aiming to optimize AI algorithms for depression detection. A ~~pivotal~~ aspect of the study lies in its cross-linguistic analysis, where the dataset is translated into Romanian, and the AI model is trained and evaluated on this multilingual data, offering insights into the model's performance in different linguistic contexts.

investigate

Through meticulous analysis and experimentation, this study presents valuable contributions to the field of depression detection, highlighting the importance of linguistic diversity in AI-based approaches. Ultimately, the findings pave the way for more effective and culturally inclusive depression identification tools, with implications for global mental health initiatives.

Contents

Chapter 1

Introduction

1.1 Understanding Depression: Statistics and Insights

Depression stands as a prevalent mental health affliction with profound impacts on both psychological and physical well-being. Characterized by a disinterest in routine activities, sleep disturbances, anhedonia, and in severe cases, suicidal ideation [Cui et al., 2015], it has become a pervasive chronic ailment across global societies, disrupting functionality, engendering despondency, and diminishing life quality. Furthermore, individuals grappling with major depressive disorder face heightened susceptibility to cardiovascular ailments, suboptimal treatment outcomes, and elevated rates of morbidity and mortality [Seligman and Nemeroff, 2015, Luo et al., 2018].

The World Health Organization (WHO) identifies depression as the primary contributor to global disability, affecting over 300 million individuals worldwide [Smith and De Torres, 2014]. Particularly alarming is the revelation that adolescents with severe depression are 30 times more prone to suicide [Stringaris, 2017]. Despite its recognized significance as a global health challenge, the intricate mechanisms underlying depression's etiology remain inadequately elucidated, albeit cultural, psychological, and biological factors are acknowledged as contributors [Gross, 2014, Ménard et al., 2016].

The Global Burden of Disease (GBD) study [Liu et al., 2020] offers comprehensive insights into various ailments across 195 countries, including depression. Divided into dysthymia and major depressive disorder categories, the GBD database from 1990 to 2017 furnishes valuable data for understanding depression's prevalence trends globally. Major depressive disorder emerges as a predominant form of depression, posing a significant burden on global health, with projections indicating it may become the leading cause of disability by 2030. Moreover, while dysthymia rates decreased in some regions, it remains a concern, particularly in the United States.

Identifying underlying causes and risk factors for depression, including genetic predisposition, demographic factors, unhealthy lifestyles, and comorbidities such as stroke, cancer, and AIDS, underscores the need for multifaceted interventions and targeted policies. Governments in countries with high depression rates are urged to prioritize research, promote healthy lifestyles, and ensure comprehensive care for individuals with predisposing conditions. However, the study acknowledges limitations in data analysis, advocating for future research to delve deeper into regional risk factors and guide tailored policy interventions for effective depression control globally, As seen in Figure 4.1 [Liu et al., 2020], which evaluated the worldwide burden of depression using the estimated annual percentage change (EAPC) and age-standardized incidence rate (ASR).

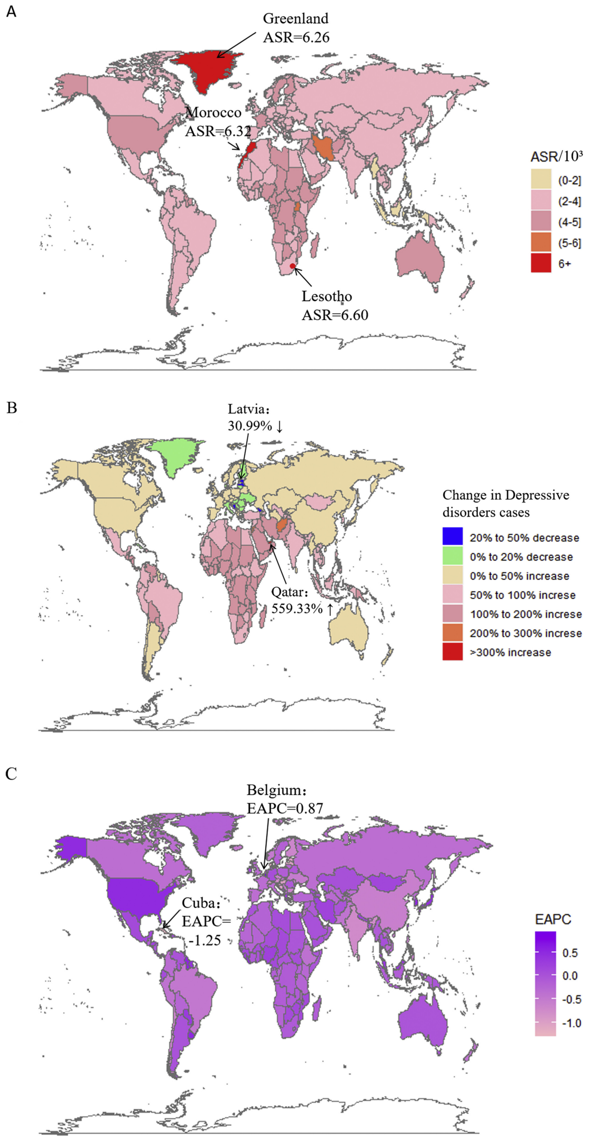


Figure 1.1: Global depression statistics comparison between 1990 and 2017 [Liu et al., 2020]

1.2 Objective: Tool for Depression Detection and Cross-Linguistic Evaluation

The objective of our scientific study is to leverage machine learning techniques to develop a robust tool for identifying depression through textual analysis. By harnessing the power of natural language processing (NLP) and artificial intelligence (AI), we aim to create a fast and reliable tool capable of detecting signs of depression in text-based communications.

The primary goal of this endeavor is twofold. Firstly, we seek to provide a timely and accessible means 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, our tool aims to offer an initial assessment of an individual's mental well-being. This proactive approach can facilitate early intervention and support, potentially mitigating the onset of more severe depressive symptoms and their associated consequences.

Secondly, we aim to evaluate the accuracy and efficacy of our machine learning model in a cross-linguistic context. To achieve this, we will translate our dataset from English to Romanian and assess the performance of the model on both language versions. This comparative analysis will enable us to ascertain the generalizability and robustness of our tool across different languages and cultural contexts.

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

Chapter 2

Navigating Textual Data: Dataset Insights and Preprocessing Techniques

2.1 Dataset Overview: Sourcing Mental Health Discussions from Subreddits

Our investigation relies on a carefully compiled dataset [Shinde, 2022], intentionally crafted to propel advancements in mental health classification research . Gathered through web scraping techniques from diverse Subreddits, this dataset encapsulates a broad range of discussions and viewpoints on mental health topics . The fundamental aim of creating this dataset was to facilitate a detailed examination of textual patterns indicative of depression's presence or absence in individuals, as inferred from their online conversations .

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

Comprising 7,650 unique entries, the dataset represents a rich tapestry of textual data. Each entry is meticulously annotated with an is_depression label, distinguishing between texts that signify 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 [Shinde, 2022].

A noteworthy aspect of the dataset is its well-balanced nature, with 3,900 entries labeled as non-depression ('0') and 3,831 entries indicating depression ('1'). This balance is instrumental in avoiding bias in the predictive modeling process, ensuring that the resulting classification model is both fair and accurate. By maintaining an almost equal distribution between the two categories, the dataset provides a solid foundation for developing robust algorithms capable of detecting signs of depression in textual data.

Given the complexities and nuances of natural language, the raw data underwent a comprehensive cleaning process using multiple Natural Language Processing (NLP) techniques. This preprocessing phase was crucial for eliminating noise, such as irrelevant characters, web links, and non-English words, thereby refining the dataset for analysis. The cleaning process also involved normalizing the text to ensure consistency across the dataset, facilitating more effective data analysis and model training [Shinde, 2022].

In summary, the dataset presents a comprehensive and balanced collection of textual data aimed at enhancing our understanding and classification of mental health states, specifically depression, through the lens of online discourse. The careful curation, cleaning, and balancing of the data underscore the rigor and thoughtfulness applied in preparing this dataset for research purposes. This foundation sets the stage for applying advanced NLP techniques and machine learning models to unravel the complexities of mental health classification based on textual analysis.

2.2 Leveraging LIWC-22 for In-depth Textual Analysis

In the realm of textual data analysis, especially within the context of psychological research, the tool we choose to process and interpret the data is as critical as the data itself. For this reason, our exploration of the dataset employs the latest version of a highly acclaimed text analysis software, LIWC-22 (Linguistic Inquiry and Word Count). This tool represents the culmination of decades of research and development in the field of computational linguistics and psychology, designed to uncover the intricate ways in which language reflects underlying psychological states.

LIWC-22 stands on the shoulders of giants, tracing its intellectual heritage back to early pioneers who first posited that the words we use in everyday communication are windows into our inner lives—revealing our thoughts, feelings, social relationships, and even our personalities. The tool is the product of a concerted effort to harness the power of computational methods to analyze language systematically, overcoming the complexities that early computer-based text analysis methods encountered [Boyd et al., 2022].

With LIWC-22, researchers have at their disposal a sophisticated software tool that not only builds upon the foundation laid by 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 unprecedented depth and precision. Whether one is interested in exploring the nuances of emotional expression, social connectivity, cognitive processes, or any other psychological dimension manifest in text, LIWC-22 offers a robust and flexible platform for investigation.

In this section, we will explore the specific features of LIWC-22 that make it an invaluable tool for our research purposes, including its methodological underpinnings, its psychometric properties, and the ways in which it allows us to parse the subtle linguistic cues that signal varying psychological states. Through this exploration, readers will gain insight into the sophisticated interplay between language and psychology that LIWC-22 helps to elucidate, setting the stage for a deeper understanding of the dataset and the insights it holds.

2.2.1 The Processing Capabilities of LIWC-22

The Linguistic Inquiry and Word Count (LIWC-22) tool stands as a cutting-edge solution for processing and analyzing textual data within the domain of psychosocial research. This sophisticated software, coupled with its comprehensive dictionary, bridges the gap between linguistic constructs and psychological theories, offering unparalleled insights into the psychosocial dimensions of language. Through a detailed overview of its primary and companion processing modules, we delve into how LIWC-22 serves as an indispensable tool for researchers aiming to uncover the psychological underpinnings of text [Boyd et al., 2022].

Upon analyzing texts, LIWC-22 quantitatively evaluates the language used against its expansive dictionary, calculating the percentage of words within each text that align with specific psychosocial categories. This process yields detailed metrics on the linguistic dimensions of the analyzed texts, which can be exported in various formats for further analysis.

Beyond its core functionality, LIWC-22 introduces several companion processing modules that enhance its analytical capabilities:

- Dictionary Workbench: Simplifying the creation of custom dictionaries, this module offers a user-friendly interface with built-in error checking. It facilitates the evaluation of custom dictionaries' psychometric properties, ensuring their effectiveness in research contexts [Boyd et al., 2022].

- Word Frequencies and Word Clouds: These features assist in identifying the most common words within a dataset, providing visual word clouds for intuitive analysis of text samples [Boyd et al., 2022].

- Topic Modeling with the Meaning Extraction Method: LIWC-22 incorporates the Meaning Extraction Method (MEM) for topic modeling, enabling researchers to uncover dominant themes and meanings within their datasets through factor analysis [Boyd et al., 2022].

- Narrative Arc: This innovative module evaluates texts for narrative structures, offering insights into storytelling elements such as staging, plot progression, and cognitive tension [Boyd et al., 2022].

- Language Style Matching (LSM): LSM analyzes the stylistic similarities between texts, offering metrics for comparing language use in various contexts, from individual communications to group dynamics [Boyd et al., 2022].

- Contextualizer: Understanding the context of word use is vital. This module extracts words along with their surrounding text, allowing for a deeper examination of linguistic usage and implications [Boyd et al., 2022].

- Case Studies: Tailored for in-depth analysis of individual texts, this module aggregates LIWC-22's capabilities to facilitate comprehensive study of specific documents or transcripts [Boyd et al., 2022].

- Prepare Transcripts: Aiding in the preparation of conversation transcripts for analysis, this module streamlines the cleaning process, ensuring texts are optimized for LIWC analysis [Boyd et al., 2022].

Together, these modules position LIWC-22 as a versatile and powerful tool for linguistic and psychological research, offering novel ways to explore and interpret the complex interplay between language and psychosocial processes.

2.2.2 The Evolution and Architecture of the LIWC-22 Dictionary

The LIWC-22 Dictionary is the linchpin of the Linguistic Inquiry and Word Count (LIWC) system, embodying the fusion of linguistic constructs with psychosocial theories through an extensive lexicon. This core component, comprising over 12,000 words, word stems, phrases, and select emoticons, is meticulously organized into categories and subcategories designed to capture a wide array of psychosocial constructs. This arrangement allows for a nuanced analysis of text, offering insights into the psychological state, social relationships, and cognitive processes of individuals based on their word usage.

Central to the LIWC-22 Dictionary's design is its 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 [Boyd et al., 2022].

The development of the LIWC-22 Dictionary represents a significant evolution from its predecessors, incorporating advances in computational linguistics and psychological research. The creation process involved multiple phases:

- Word Collection: Leveraging the foundation of the LIWC2015 dictionary, new words were generated for each category through a combination of expert input and comprehensive literature review [Boyd et al., 2022].
- Judge Rating Phase: Words were qualitatively assessed by a panel of judges for their fit within each category, with disagreements resolved through in-depth analysis and consensus. Base Rate Analyses: Utilizing the Meaning Extraction Helper (MEH) tool, the frequency of dictionary words in a diverse corpus was evaluated to ensure relevance and applicability across various text samples [Boyd et al., 2022].
- Candidate Word List Generation: Through statistical analysis and expert review, candidate words were identified for potential inclusion in the dictionary, ensuring a broad and relevant lexicon [Boyd et al., 2022].
- Psychometric Evaluation: Each category underwent rigorous testing for internal consistency, with adjustments made to optimize the dictionary's psychometric properties [Boyd et al., 2022].
- Refinement Phase: The entire process was iteratively refined to address any oversights and enhance the dictionary's accuracy and reliability [Boyd et al., 2022].
- Addition of Summary Variables: New summary variables were introduced to provide additional analytical dimensions, based on cutting-edge research [Boyd et al., 2022].

The LIWC-22 Dictionary has been significantly expanded to include not only traditional words but also numbers, punctuation, short phrases, and regular expressions. This expansion allows for the analysis of modern, informal communication styles found on social media and text messaging, incorporating ”netspeak” and emoticons for a more comprehensive understanding of digital communication.

The dictionary's evolution reflects a balance between expert human judgment and sophisticated computational models, ensuring that LIWC-22 remains at the forefront of text analysis technology. With each iteration, LIWC has adapted to the changing landscape of language use, incorporating new categories and adjusting existing ones to better capture the psychological significance of language [Boyd et al., 2022].

In summary, the LIWC-22 Dictionary's development and structure are a testament to the interdisciplinary collaboration between linguistics and psychology. Its comprehensive and adaptable design makes it an invaluable tool for researchers and practitioners seeking to understand the deep psychosocial underpinnings of language use.

2.2.3 The Psychometric Rigor of LIWC-22: Establishing Reliability and Validity

The development of the Linguistic Inquiry and Word Count (LIWC-22) tool has consistently prioritized the establishment of a scientifically robust system, focusing on both reliability and validity. This commitment has guided each iteration of LIWC, with the aim of adapting to the dynamic nature of language use and leveraging the expanding horizons of text-based data science. LIWC-22 represents the culmination of these efforts, integrating modernized dictionaries with cutting-edge data analytics to offer a highly validated tool for text analysis.

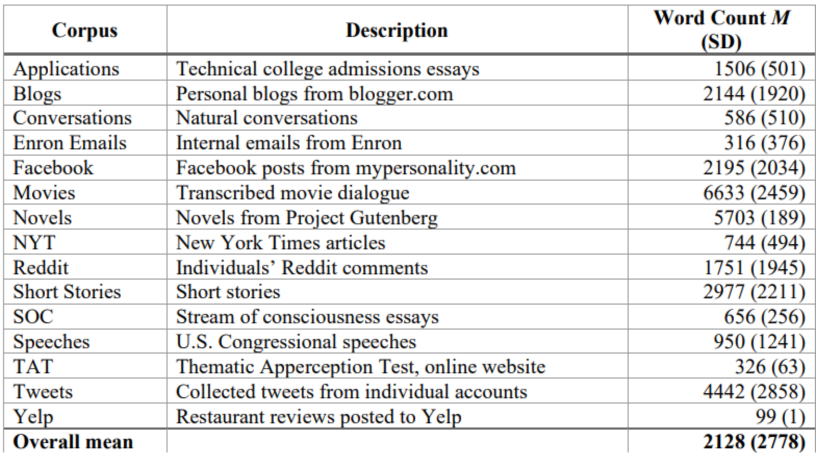
At the heart of LIWC-22's psychometric establishment is the "Test Kitchen" corpus [Figure 2.1], a meticulously curated collection of English language samples drawn from a wide spectrum of sources. This corpus serves dual purposes: it is instrumental in the selection of words for the LIWC-22 dictionary and plays a crucial role in assessing the dictionary's reliability and validity. The breadth and diversity of the Test Kitchen corpus [Figure 2.1] ensure that LIWC-22's analyses are grounded in a realistic representation of language use across various contexts [Boyd et al., 2022].

To capture the multifaceted nature of language, the Test Kitchen corpus [Figure 2.1] was assembled from 15 distinct English language data sets, encompassing a wide range of communication forms, from blogs and emails to social media posts and movie dialogues. This comprehensive corpus consists of 15,000 texts, with each text sample reflecting the unique linguistic style of its author(s). The selection process for these samples was designed to include a diverse representation of texts, ensuring a broad coverage of language use in daily life.

The construction of this corpus involved selecting 1,000 text samples from each of the 15 sources, with each text containing at least 100 words. For longer texts, a specific algorithm was employed to extract a continuous segment of 10,000 words, ensuring a manageable and consistent analysis size. In total, the Test Kitchen corpus [Figure 2.1] encompasses over 31 million words, providing a robust foundation for the validation and refinement of the LIWC-22 dictionary [Boyd et al., 2022].

Given the sensitivity and proprietary nature of some of the data sources, the Test Kitchen corpus, while invaluable for the development and testing of LIWC-22, cannot be made publicly available. This restriction underscores the careful consideration given to privacy and ethical research practices in the compilation and use of the corpus. Nevertheless, the corpus's diverse and extensive dataset has been crucial in fine-tuning LIWC-22's dictionaries to reflect genuine language usage patterns.

The meticulous construction of the Test Kitchen corpus and its application in developing LIWC-22 illustrate the comprehensive approach taken to ensure the tool's psychometric integrity. By grounding the dictionary in a wide-ranging and representative sample of English language use, LIWC-22 stands as a testament to the evolving field of text analysis, offering researchers a reliable and valid instrument for exploring the depths of linguistic and psychosocial phenomena [Boyd et al., 2022].



Corpus	Description	Word Count *M* (SD)
Applications	Technical college admissions essays	1506 (501)
Blogs	Personal blogs from blogger.com	2144 (1920)
Conversations	Natural conversations	586 (510)
Enron Emails	Internal emails from Enron	316 (376)
Facebook	Facebook posts from mypersonality.com	2195 (2034)
Movies	Transcribed movie dialogue	6633 (2459)
Novels	Novels from Project Gutenberg	5703 (189)
NYT	New York Times articles	744 (494)
Reddit	Individuals' Reddit comments	1751 (1945)
Short Stories	Short stories	2977 (2211)
SOC	Stream of consciousness essays	656 (256)
Speeches	U.S. Congressional speeches	950 (1241)
TAT	Thematic Apperception Test, online website	326 (63)
Tweets	Collected tweets from individual accounts	4442 (2858)
Yelp	Restaurant reviews posted to Yelp	99 (1)
Overall mean		**2128 (2778)**

Figure 2.1: The test Kitchen Corpus of 31 Million Words [Boyd et al., 2022]

2.2.4 Challenges and Methodologies in Assessing LIWC-22's Psychometrics

The process of quantifying the reliability and validity of text analysis tools like LIWC-22 presents unique challenges, diverging significantly from the conventional approaches used in psychological assessments. The inherent differences between verbal behavior and structured questionnaire responses necessitate a nuanced approach to evaluating the psychometric properties of LIWC categories.

Unlike self-report questionnaires that gauge a construct like anger through multiple, similar questions to ensure internal consistency, natural language does not conform to such repetitive patterns. In real-world communication—be it a social media post, an essay, or a conversation—individuals express a thought and then naturally progress to the next, without the redundant expression of the same idea. This characteristic of verbal expression implies that the psychometric standards applied to language-based analyses must be recalibrated to account for the unique dynamics of verbal behavior [Boyd et al., 2022].

The evaluation of LIWC-22's reliability involves an innovative adaptation to the language's non-repetitive nature. Taking the LIWC-22 Anger scale as an instance, the scale encompasses 181 words and phrases associated with anger. Theoretically, the usage of one anger-related word in a text should correlate with the usage of other anger-related words within the same text. By analyzing how each of these words is employed across a selection of texts and calculating the intercorrelations among these word usages, LIWC-22's approach to determining internal consistency emerges[Boyd et al., 2022].

Validating the numerous LIWC dimensions poses a significant and complex challenge. By their nature, LIWC's content categories appear to be directly relevant or face valid. Yet, the deeper question lies in determining the extent to which both personal and social psychological processes are mirrored in the use of language. For instance, the implications of using words related to "affiliation" at a high frequency raise questions about the user's social connections and needs. Are individuals using these words seeking more social interaction, or do they reflect a person's existing strong social ties? Additionally, it is important to consider whether the frequency of such language usage offers insights into or predictions about someone's social relationships and needs.

To compute these metrics, LIWC-22 employs two statistical methods: the Cronbach's alpha (α) for continuous data, based on the percentage of total words, and the Kuder–Richardson Formula 20 (KR-20) for binary data [Kuder and Richardson, 1937], indicating the presence or absence of words. These methods yield insights into the internal consistency of the LIWC-22 categories, adapting traditional psychometric calculations to the context of language analysis [Boyd et al., 2022].

The application of Cronbach's alpha in the context of LIWC-22 encounters a significant hurdle due to the variable base rates of word usage within language categories. This variability can lead to underestimations of reliability when using traditional methods. Conversely, the Kuder–Richardson Formula 20 offers a more accurate reflection of a category's internal consistency by accommodating the binary nature of word presence, thus providing a "truer" approximation of reliability in language analysis.

The volume of research at the intersection of text analysis and psychosocial processes is vast, with over 2,400 studies cited in 2021 alone that utilized LIWC for text analysis. Findings from these studies, including those from the developers' own laboratories, reveal correlations between the affect or emotion categories detected by LIWC in texts and the authors' self-reported feelings. These correlations, although modest, underscore the tool's capability to capture psychological dynamics to a certain extent. Higher correlations are observed when comparing judges' ratings of writing samples with LIWC scores, suggesting a somewhat consistent external validation of LIWC's analytical output[Boyd et al., 2022].

The methodologies adopted for assessing the reliability and validity of LIWC-22 underscore the tool's sophisticated approach to text analysis. By carefully navigating the intricacies of natural language and employing tailored statistical methods, LIWC-22 achieves a nuanced and psychometrically sound analysis of verbal behavior. This approach not only highlights the challenges inherent in language-based psychometrics but also showcases LIWC-22's commitment to providing reliable and valid insights into the psychological underpinnings of text.

2.3 Tokenization for LIWC-22 Compatibility

We shall delve into the tokenization method utilized in our study, designed to preprocess text data to ensure it aligns with the requirements of the Linguistic Inquiry and Word Count (LIWC-22) tool [Boyd et al., 2022]. LIWC requires input texts to be broken down into tokens, a process that can significantly influence the accuracy and reliability of the linguistic analysis. Our chosen methodology leverages the strengths of BERT's tokenization system, as described in the referenced study [Devlin et al., 2018].

BERT, or Bidirectional Encoder Representations from Transformers, introduces a sophisticated approach to tokenization that we adapted for our needs. BERT's tokenization algorithm is based on WordPiece [Wu et al., 2016], handling the input text by initially breaking it down into tokens, which are then further divided into sub-tokens. This mechanism allows for a fine-grained understanding of language, capturing nuances by analyzing tokens in the context of their surrounding text.

The BERT model is pre-trained on a vast corpus of text, allowing it to understand a wide range of linguistic nuances. Its bidirectional nature means that each token is influenced by the tokens that come before and after it, providing a rich context for each word. This context is crucial for accurate linguistic analysis, as the meaning of a word can change significantly depending on its context [Devlin et al., 2018]. For example, in the sentence "I read that book," the tokenization process needs to understand whether "read" is in the past or present tense, which has implications for the psychological constructs that LIWC-22 might extract.

By using BERT's tokenization, we ensured that our model could handle the complexity of the text data typically found on platforms like Reddit. This is particularly important for detecting signs of depression, where context can change the sentiment or meaning of a word. Furthermore, the sub-token approach allowed us to maintain the granularity needed for LIWC-22 [Boyd et al., 2022], which often relies on the detection of specific words and categories relevant to psychological states.

Chapter 3

Model Selection and Hyperparameter Tuning: Optimizing AI for Depression Detection

In the quest to develop an effective AI system for detecting depression from textual data, the choice of the right model and the fine-tuning of its parameters emerge as critical steps. This chapter delves into the intricate process of selecting Random Forest as the preferred model for our task. Known for its robustness and ability to handle complex datasets, Random Forest stands out as a powerful tool in the landscape of machine learning algorithms. However, the journey from selection to optimization is nuanced, involving a series of strategic decisions aimed at enhancing the model's performance.

3.1 The Strategic Choice of Random Forest for Depression Detection

In the ever-expanding realm of machine learning, selecting the most appropriate algorithm is paramount to the success of any predictive modeling task. This is particularly true in the domain of depression detection, where the complexity and variability of the data demand an approach that is not only accurate but also robust and interpretable. Drawing upon the findings of a comprehensive study that evaluated twelve distinct machine learning algorithms across seven datasets[Siraj-Ud-Doulah and Islam, 2023], we anchored our decision to employ Random Forest (RF) as the cornerstone of our analysis.

The study [Siraj-Ud-Doulah and Islam, 2023] in question meticulously 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 emerged as the clear frontrunner, exhibiting 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 [Siraj-Ud-Doulah and Islam, 2023] also highlighted the significance of the kappa statistic and Root Mean Square Error (RMSE) as vital factors in assessing model performance, further validating the robustness of Random Forest in handling diverse and complex datasets. Inspired by these compelling insights, and in alignment with the study's conclusion, our selection of Random Forest is underpinned by its demonstrated efficacy across multiple evaluative dimensions.

The datasets utilized for the comparative study are varied, each with its unique characteristics and relevance to different classification tasks:

- Breast Cancer Wisconsin (Original): This dataset contains 11 attributes and is used for binary classification (two classes) with 699 instances. It does include missing values, which would require additional preprocessing steps.

- Statlog (Vehicle Silhouettes): Comprising 19 attributes over 846 instances, this dataset is for multiclass classification with four distinct classes and has no missing values.

- Vertebral Column: With 7 attributes and 310 instances, this dataset is also used for multiclass classification, distinguishing among three classes, without any missing values.

- Breast Tissue: This dataset has 10 attributes across 106 instances and is used for a more complex multiclass classification task with six classes, also free of missing values.

- Contraceptive Method Choice: It includes 10 attributes and a larger number of instances at 1473. It's structured for multiclass classification into three classes, and there are no missing values.

- Image Segmentation: This is a sizable dataset with 20 attributes and 2310 instances for multiclass classification involving seven classes, and it contains no missing values.

- Artificial Characters: The largest among the datasets listed, it boasts 8 attributes across a substantial 10218 instances. It's designed for a multiclass classification with ten classes, and like most others here, it lacks missing values.

In the context of our study focused on depression detection, our model resembles the Breast Cancer Wisconsin dataset, because we are also tackling a binary classification problem. However, our model differentiates itself with a higher dimensionality, processing 64 input attributes, which poses a greater complexity in feature representation and selection. (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 topped the charts with the highest kappa value of 95.03%, indicating a strong agreement between the predictions and the actual classifications. Precision with RF was outstanding as well, hitting a high of 98%, while its recall was nearly as impressive at 97.9%, underscoring its ability to identify most of the positive cases.

Across the rest of the datasets analyzed in the study [Siraj-Ud-Doulah and Islam, 2023], Random Forest (RF) consistently delivered standout performance. Its F-measure and Matthew's Correlation Coefficient (MCC) values were notably high, often outperforming other algorithms. For instance, RF attained an accuracy of 98.48%, kappa value of 98.23%, and precision and recall rates both at 98.5% on certain datasets, alongside an exceptional specificity of up to 99.7

While K-NN and Logistic Regression (LR) also demonstrated strong performances in certain cases, with K-NN leading in precision and recall in the Breast Tissue dataset and LR excelling with the highest MCC values for the Vehicle and Vertebral Column datasets, RF's overall dominance was clear. RF's ability to achieve the lowest error rates, coupled with the lowest root mean square error in the majority of datasets, further confirms its robustness and reliability as an algorithm for complex predictive tasks, including depression detection.

In summary, the numerical evidence from the study [Siraj-Ud-Doulah and Islam, 2023] underlines RF's superior ability to handle complex predictive tasks, making it the algorithm of choice for our model aimed at accurately detecting depressive patterns within textual data.

Chapter 4

Cross-Linguistic Analysis: Translating and Training the AI Model for Multilingual Depression Detection

In the ever-evolving field of Artificial Intelligence (AI) and Natural Language Processing (NLP), the ability to accurately detect signs of depression across different languages is both a challenge and a necessity. Multilingual depression detection hinges on the capability of AI models to understand and analyze text beyond the confines of a single language. This section delves into the critical process of translating English text into Romanian, a step essential for training our AI model to recognize depressive patterns in a multilingual context.

We will discuss the selection criteria and the impact of utilizing a specific Translation API to bridge the language gap, thus enabling our model to process and interpret Romanian text with the same level of proficiency as English. By incorporating these translation mechanisms, we aim to enhance the model's sensitivity and accuracy in identifying depression indicators across diverse linguistic landscapes.

4.1 Comparative Analysis of Translation APIs and Selection Rationale for Yandex

In our effort to refine our multilingual depression detection model, we referenced a detailed study that assessed the efficiency, accuracy, and security of various Translation APIs [Rashmi et al., 2020]. This comparative analysis served as the foundation for selecting the most suitable API for our application, which required the translation of text from English to Romanian among other language pairs. The study meticulously compared several leading Translation APIs, including Google API, Microsoft, Systran.io, MyMemory, and Yandex, focusing on their performance in terms of speed, accuracy, security, and the breadth of language support.

- **Google API** is widely recognized for its impressive language support, capable of translating content across more than 100 languages. This extensive reach makes it a versatile tool for global communication and content translation. Its reputation and prevalence in the market are testaments to its utility and user-friendly interface. Additionally, it's worth noting that while Google API is commendable, it is not a free service, which may affect its accessibility for some users.[Rashmi et al., 2020].

- **Microsoft's Translation API** is lauded for its quality and security, offering translations among 60+ languages. It stands out for its emphasis on accuracy and stringent security protocols, although its language support is less extensive than Google's [Rashmi et al., 2020].

- **Systran.io** boasts a high accuracy rate of 99%, albeit with limitations in recognizing slang, nuances, and culturally relevant phrases. Its security is commendable, positioning it as a reliable choice for many applications [Rashmi et al., 2020].

- **MyMemory** excels in translation speed but experiences the highest latency among the APIs evaluated. While it supports translations between 80+ languages, the absence of training data for certain language combinations limits its effectiveness. Nonetheless, its security is robust [Rashmi et al., 2020].

- **Yandex API**, with support for 90+ languages, stands out for its balance of translation accuracy and lower latency compared to its counterparts. Despite its efficiency and broad language coverage, its security features are not optimal for translating confidential documents [Rashmi et al., 2020].

The conclusion from this study illuminated the strengths and weaknesses of each API, guiding our choice towards Yandex API for our multilingual depression detection model. Yandex was selected due to its free access, lower latency, and reliable accuracy across complex language pairs, making it an ideal tool for everyday translations where security is not the paramount concern [Rashmi et al., 2020]. This choice aligns with our objective of enhancing accessibility and efficiency in depression detection across multiple languages without the need for extensive resources or development time.

Further bolstering our decision to incorporate Yandex into our multilingual depression detection framework is another rigorous study that provides a nuanced error analysis of Yandex's translations. Notably, Yandex's performance, depicted in the accompanying graph, indicates a relatively uniform distribution of errors across multiple categories [Cambedda et al., 2021]. This suggests that while Yandex does have areas that require attention, such as Lexis, Syntax, and Article Usage, it generally maintains the core meaning of the translated text. This is critical for our model, which relies on the preservation of semantic content to accurately detect depressive indicators in text.

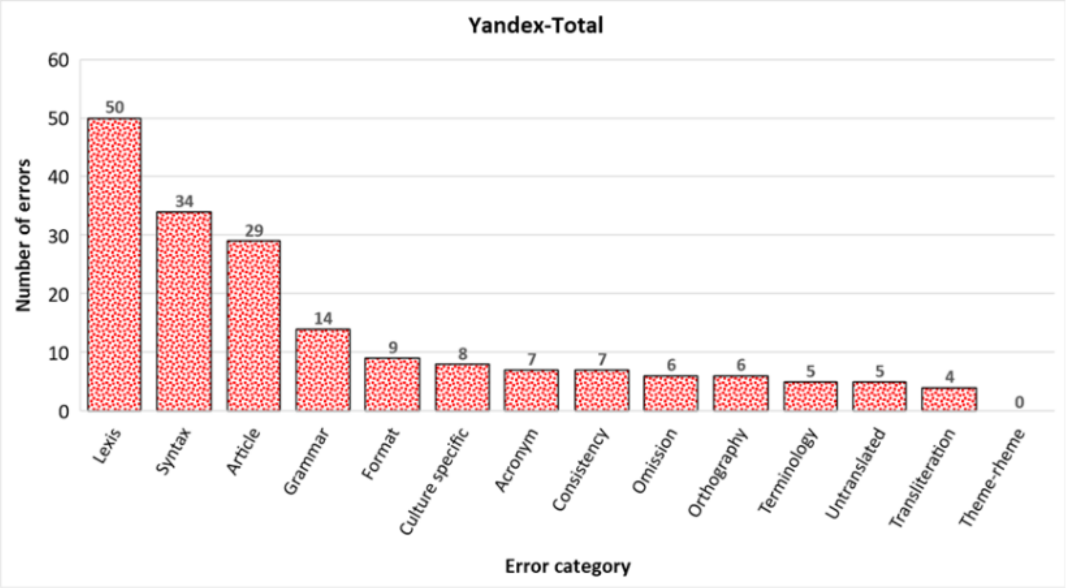


Figure 4.1: Yandex's translation performance [Cambedda et al., 2021]

The Lexis category, in particular, displays the highest number of errors, signaling a need for further investigation to understand whether these issues arise from the nature of the texts or from inherent challenges within the translation tool. However, it is encouraging to note that error categories such as Grammar, Format, Culture-Specific References, Acronym, and Consistency show significantly fewer errors [Cambedda et al., 2021]. These categories are essential for maintaining the integrity of meaning, which reaffirms our choice of Yandex for texts where nuanced meaning is less likely to affect the detection of depression indicators.

Further examination of the study indicates that Yandex demonstrates a more adept handling of common language texts as opposed to those with specialized jargon, registering fewer errors in translations of texts with general vernacular [Cambedda et al., 2021]. Considering our dataset comprises of Reddit posts, which are typically phrased in everyday language, this finding is particularly relevant. The study also highlights that, with the exception of Article Usage, the disparity in error rates between different text types is negligible, suggesting that Yandex can reliably manage the conversational and informal style characteristic of Reddit communications.

The study's findings underscore Yandex's capability to offer a satisfactory level of precision and effectiveness for our dataset. Given that Reddit posts are less formal, Yandex's translation services appear well-suited for our project's requirements. The platform's proficiency in handling everyday language makes it an ideal candidate for our depression detection model's multilingual component. It provides us with a valuable tool for expanding our model's reach, ensuring that the essence of the messages is captured, which is essential for accurate sentiment analysis, even if minute linguistic details may not be perfectly preserved.

4.2 Adjusting to Yandex API's Policy Shift: Navigating New Constraints

In the pursuit of refining our multilingual depression detection model, we had initially recognized the Yandex API as a superior option, particularly for its cost-effectiveness, as it was freely accessible at the time of study [Rashmi et al., 2020]. This advantage aligned seamlessly with our objectives, allowing us to leverage its translation capabilities without financial constraints, facilitating broader research and application development.

However, since the publication of [Rashmi et al., 2020], Yandex's policy landscape has undergone significant changes. The API, once celebrated for its complimentary access, has shifted to a model that requires users to possess a registered and legally recognized company to utilize its services. This pivot in policy necessitates a reassessment of our tool selection criteria and the potential impact on our project's scope and resource allocation.

The requirement of company registration introduces a layer of complexity, potentially limiting the accessibility of Yandex API for independent researchers, small teams, or educational institutions that may lack formal corporate structures. It also prompts us to consider the legal and administrative overhead that accompanies the establishment of a formal entity, which may not be viable or desirable for all projects.

In light of these new stipulations, our commitment to developing an effective and accessible multilingual depression detection model remains unwavering. As such, we are prompted to explore alternative strategies.

4.3 Transitioning to googletrans for Multilingual Support

After careful consideration of the new constraints imposed by Yandex API, our team has made a strategic pivot to integrate the googletrans library [Suhun, 2020] into our multilingual depression detection model. googletrans presents itself as an appealing alternative, offering a free and unlimited Python library that interfaces with the Google Translate API. This library appears particularly advantageous for our requirements, as it is not only accessible without cost but also does not require the bureaucratic process of company registration that Yandex now demands.

The googletrans library [Suhun, 2020] boasts impressive features that are well-suited to our project's needs. It is recognized for its speed and reliability, as it operates on the same servers as translate.google.com. The library supports auto language detection, facilitating the identification and translation of a wide array of languages without prior specification. Additionally, it provides the capability for bulk translations, which is invaluable when processing large datasets typically found in NLP tasks.

While googletrans [Suhun, 2020] has an impressive array of features, it is also important to acknowledge the library's usage notes. The 15,000-character limit per text may require segmentation of longer passages, and the inherent instability of web-based translation services means that we should proceed with a level of flexibility regarding the library's reliability. The developers themselves suggest opting for the official Google Translate API for critical applications where stability is paramount. Furthermore, potential HTTP errors could indicate temporary bans by Google, necessitating monitoring and management of our API usage to prevent disruption.

Despite these considerations, the googletrans library's free and robust nature makes it an excellent fit for our project in its current stage. It allows us to continue advancing our multilingual depression detection capabilities while adhering to our resource constraints. In this section, we will explore the integration process, address the library's limitations with strategic solutions, and outline how we will ensure the model's performance remains high even with the switch to a different translation tool.

4.4 Optimizing Tokenization for Romanian: Preserving Preprocessing Uniformity Across Languages

When constructing a multilingual depression detection model, the choice of tokenizer is pivotal for accurately interpreting the linguistic nuances of each language. Our study contrasts the tokenization techniques of English with those adapted for Romanian, based on the comprehensive study of Romanian BERT [Dumitrescu et al., 2020]. Tokenization, the breaking down of text into its constituent parts or tokens, serves as the foundation for pre-processing text data.

BERT's tokenizer for English text employs a WordPiece model[Wu et al., 2016], which excels at parsing English sentences into a sensible sequence of sub-tokens.

This tokenizer, while effective, is optimized for the linguistic patterns inherent in the English language, which differ significantly from Romanian. Romanian Bert [Dumitrescu et al., 2020] tailors the tokenization process to the Romanian language's unique characteristics.

The comparative advantage of the Romanian-specific tokenizer over a more general multilingual BERT (M-BERT) model is evident in its proficiency at tokenization, crucial for any NLP task. In the context of our model's use case, it was shown that the Romanian BERT tokenizer could break down words into approximately 1.4 tokens on average, while M-BERT reached up to 2 tokens per word for the cased vocabulary. Additionally, the incidence of unknown tokens was drastically reduced by an order of magnitude with the Romanian BERT tokenizer [Dumitrescu et al., 2020].

This optimized tokenization not only enhances the accuracy of linguistic analysis but also improves the model's ability to interpret the text in a manner that aligns with LIWC-22's [Boyd et al., 2022] requirements. With better tokenization, the nuances of depression indicators in the text are more likely to be captured, regardless of linguistic differences. This tailoring becomes all the more crucial when the LIWC tool is applied, as it relies heavily on token recognition to categorize and quantify various linguistic and psychological components within the text.

The study's conclusion [Dumitrescu et al., 2020] underscored the superiority of the Romanian BERT tokenizer for the Romanian language, illustrating the benefits of customizing NLP tools to accommodate the linguistic intricacies of specific languages. This aligns with our goal of achieving high fidelity in detecting depressive markers within text, bolstering the model's sensitivity and precision, especially in a language-sensitive context like mental health assessment.

In essence, the custom Romanian tokenizer is not only better suited for handling Romanian text but also exemplifies the importance of language-specific NLP tools in enhancing the performance of models on tasks such as depression detection from social media text.

Chapter 5

Conclusions

Concluzii ...

Bibliography

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

[Cambedda et al., 2021] Cambedda, G., Di Nunzio, G. M., and Nosilia, V. (2021). A study on automatic machine translation tools: A comparative error analysis between deepl and yandex for russian-italian medical translation. *Umanistica Digitale*, (10):139–163.

[Cui et al., 2015] Cui, R. et al. (2015). A systematic review of depression. *Curr Neuropharmacol*, 13(4):480.

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

[Dumitrescu et al., 2020] Dumitrescu, S. D., Avram, A.-M., and Pyysalo, S. (2020). The birth of romanian bert. *arXiv preprint arXiv:2009.08712*.

[Gross, 2014] Gross, M. (2014). Silver linings for patients with depression? *Current Biology*, 24(18):R851–R854.

[Kuder and Richardson, 1937] Kuder, G. F. and Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3):151–160.

[Liu et al., 2020] Liu, Q., He, H., Yang, J., Feng, X., Zhao, F., and Lyu, J. (2020). Changes in the global burden of depression from 1990 to 2017: Findings from the global burden of disease study. *Journal of psychiatric research*, 126:134–140.

[Luo et al., 2018] Luo, Y., Zhang, S., Zheng, R., Xu, L., and Wu, J. (2018). Effects of depression on heart rate variability in elderly patients with stable coronary artery disease. *Journal of Evidence-Based Medicine*, 11(4):242–245.

[Ménard et al., 2016] Ménard, C., Hodes, G. E., and Russo, S. J. (2016). Pathogenesis of depression: Insights from human and rodent studies. *Neuroscience*, 321:138–162.

[Rashmi et al., 2020] Rashmi, C., John, N. A., Choudhary, M., and Devraj, M. (2020). Comparison between leading apis used in translation apps. *International Journal of Advanced Research in Computer Science*, 11.

[Seligman and Nemeroff, 2015] Seligman, F. and Nemeroff, C. B. (2015). The interface of depression and cardiovascular disease: therapeutic implications. *Annals of the New York Academy of Sciences*, 1345(1):25–35.

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

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

[Smith and De Torres, 2014] Smith, K. and De Torres, I. (2014). A world of depression. *Nature*, 515(181):10–1038.

[Stringaris, 2017] Stringaris, A. (2017). What is depression?

[Suhun, 2020] 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].

[Wu et al., 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.