UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Sílvia Martins Guerra

Contextual NLP Explanations for Language Biomarker Research: Identification of Schizophrenia Traits on Social Media Posts using Multilevel Part Of Speech Feature

Belo Horizonte
2023

Sílvia Martins Guerra

Contextual NLP Explanations for Language Biomarker Research: Identification of Schizophrenia Traits on Social Media Posts using Multilevel Part Of Speech Feature

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Adriano Alonso Veloso
Co-Advisor: Cilene Aparecida Nunes Rodrigues

Belo Horizonte
2023

Resumo

Técnicas de PLN tem sido cada vez mais aplicadas à diversos cenários, se tornando especialmente valiosas na pesquisa em psicolinguística, onde surgem como uma ferramenta para gerar insights e medir de forma eficiente biomarcadores de transtornos mentais, contribuindo para diagnóstico e controle. Um grande desafio atual é criar metodologias multidisciplinares centradas na compreensão humana, que permitam aos linguistas a interpretação dos resultados e a criação de experimentos cada vez mais direcionados. Recentemente, a pesquisa de distúrbios de linguagem em pessoas com esquizofrenia têm feito progressos significativos. Nós propomos uma abordagem diferenciada para o design de features usando um método multi-nivel de POS-tagging, treinamos diversos modelos de aprendizado de máquina com diferentes grupos de features baseadas em POS-tagging, e comparamos os resultados. Além da análise da métricas de performance, demonstramos com o uso de técnicas de explicabilidade o como essa abordagem de design de features permite uma melhor exploração dos resultados, proporcionando oportunidades de análises aprofundadas e novos insights.

Palavras-chave: PLN, aprendizado de máquina, psicolinguística, computação, esquizofrenia, biomarcador, linguagem, POS-tag, explicabilidade, design de features

Abstract

The use of NLP techniques has been increasingly deployed in a wide variety of settings and has become especially valuable in physicholinguistic research, where it is emerging as a tool to find new insights and efficiently measure biomarkers for mental disorders, contributing to diagnosis and control. The need to create multidisciplinary human-centered methodologies to allow linguists to make sense of the results and design directed experiments is a pressing challenge. Recently, the research on language disorders in people with schizophrenia has been making significant progress. We proposed an approach to feature engineering using a multilevel POS tagging method, trained several machine learning models with the different levels of POS-tagging-based features, and compared their results. Beyond performance metrics analysis, we demonstrate with the use of explainability techniques how this feature design approach allows more exploration of the results and can provide valuable in-depth analysis opportunities and insights.

Keywords: NLP, machine learning, physicholinguistic , computing, schizophrenia, biomarker, language, POS-tag, explainability, feature design

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Natural language refers to human spoken, written, or sign language that evolved organically through variation, inheritance, and selection. The study of Natural Language Processing (NLP) has been ongoing for over 60 years, constantly transforming and advancing, and it has developed significantly in recent years. Nowadays, NLP-based solutions are routinely present in science, businesses, and people's lives. It is a multidisciplinary area that encompasses computer science, linguistics and artificial intelligence, and uses several computational techniques to analyze large amounts of natural language data.

Text classification is a common NLP task that aims to automatically analyze text and then assign predefined tags or categories based on its context. Rule-based approaches to text classification, by manually creating and programming a set of instructions for classification, are time-demanding, costly, and mostly lead to systems that rapidly become outdated. A more practical and efficient approach involves machine learning techniques, in which classification rules are automatically created by processing and extracting patterns from data. Moreover, this process allows new insights into the reasons for the distinction of the classes since in the patterns found by the model new information might emerge.

Machine Learning classifiers learn to make a classification based on past observations from the datasets and, for that, they benefit from having a large amount of data available. Creating a specialized corpus for specific language-related problems and domains is an essential step that might require experimentation, manual data collection, web-scrapping, text treatment, and annotation. Annotation adds value to a corpus extending the range of questions a corpus can address. The annotation of the corpus can be done by field specialists or automatically, using defined rules or even NLP models. Annotation is often achieved by a combination of approaches.

Experiments with Machine Learning classifiers on specialized corpus are a valuable resource for linguistic research, especially in the field of psycholinguistics, a transdisciplinary field that includes elements from linguistics, cognitive science, psychology, and neuroscience. Some mental disorders' fundamental symptoms include abnormalities of cognition and language, and identifying these linguistic difficulties is essential for properly characterizing the conditions. Such investigation has been essential in the understanding of Schizophrenia.

Schizophrenia is a complex neurologically condition that impairs one's capacity to manage thoughts, feelings, and interpersonal relationships. These abnormalities frequently affect capacity in the workplace, in relationships with others, and in education. Numerous grammatical deficiencies have been identified in recent studies, and there's an ongoing effort to quantify these linguistic anomalies, group them, and distinguish their presentation differences for specific communication contexts and languages.

1.1 Research Statement

This work aims at contributing to advancement in the development of reliable and automatic quantitative metrics that can significantly help to detect and treat schizophrenia. To this end, we seek to contribute to the design of a multidisciplinary research methodology applied to the study of language as a biomarker. We consider that an approach that includes the design and selection of features with different hierarchical levels of lexical specificity, machine learning methods, and model explainability techniques, can provide an innovative insightful method to analyze and measure linguistic symptoms of schizophrenia

1.2 Research Questions

This work investigates two main research questions:

- Can combined machine learning techniques be successfully applied to identify texts of people with schizophrenia, advancing further our scientific understanding of the linguistic traits of schizophrenia?

- Given three sets of features with different hierarchical levels of lexical and grammatical specificity, and a mixed set with features from all levels, which group of features can produce a model that better generalizes the results ?

1.3 Organization

The thesis is structured as follows :

- Chapter 1 - Introduction

This chapter introduces the subject of the thesis, presenting its objectives and research questions.

- Chapter 2 - Schizophrenia and Language Disorder

This chapter presents the main linguistic profile of people diagnosed with schizophrenia, an overview of recent advancements in the understanding of the field, and the related literature. The information presented in this chapter is essential to the modeling of the problem and to the discussion of the experiment results.

- Chapter 3 - NLP and Text Classification

In this chapter, we explain machine learning and natural language processing concepts that are relevant to our experiment. We describe the steps in a machine learning project and considerations on a classification task, model explainability, and text representation. The chapter's discussion also encompasses important learning methods that are fundamental to the experiment's chosen algorithms.

- Chapter 4 - Data and Algorithms

In this chapter, we describe the data source, the construction of the corpus , the feature engineering method and the models algorithms.

- Chapter 5 - Experiment

In this chapter, we describe the experiments feature selection, baseline model, model training, evaluation, explainability, and results from analysis.

- Chapter 6 - Conclusion

This chapter provides a conclusion to the presented work, its limitations, and suggestions for future explorations with the proposed approach.

Chapter 2

Schizophrenia and Language Disorder

2.1 Schizophrenia

The term schizophrenia was derived from two greek words, *schizo* and *phrene*, which respectively mean split and mind. Though the word was first used in 1908 by the swiss psychiatrist Eugen Bleuler to convey fragmented thinking as a misalignment of thinking, memory, and perception, there were already reports of mental suffering that could fit the description. What we now recognize as schizophrenia is the result of a concept that has undergone numerous changes [14]. Nevertheless, since the conceptualization of schizophrenia more than a century ago, four sub-types have been used to describe the multiplicity of the disorder, namely disorganized, catatonic, paranoid, and undifferentiated schizophrenia. The diagnosis is based on a variety of distinct behaviors and reported symptoms. Thus, two patients can be given the diagnosis of schizophrenia without sharing most of their symptoms. Unfortunately, there is no reliable biological marker evidencing an underlying biological process[50].

A person with schizophrenia at times, or consistently, interprets reality abnormally, failing to produce coherent thoughts about themselves and their surrounding reality. The presence of schizophrenia is characterized by a combination of symptoms such as hearing voices, visual hallucinations, delusions, disordered thinking, and abnormal behavior. Those symptoms can be impairing, or even disabling, for a person's personal, professional and social life. It is a lifelong condition, but there is a wide range of care options, including medication, psychoeducation, cognitive-behavioral therapy, and psychosocial rehabilitation. At least one in three people with schizophrenia that receives treatment is able to become fully functional [17].

Schizophrenia symptoms are grouped into positive, negative, and disorganized. The positive symptoms are related to changes in behavior or thoughts, such as hallucinations, delusions, and disorganized speech, while negative symptoms refer to a lack of healthy traits, like concentration, motivation, or interest in daily activities. Additionally, disorganized symptoms account for deficits in cognitive abilities, executive skills, and memory. Over time, symptoms might change in nature and degree, with periods when they get worse and times when they go away. Some symptoms could be present at all times [21].

There is no known precise cause of schizophrenia. A person may be more susceptible to developing the disorder if a combination of physical, genetic, psychological, and environmental factors are present. It is known that if a person presents a genetic predisposition, a difficult or upsetting life event could set off a psychotic episode. The reason why some people experience specific symptoms while others do not is unknown. Genetic research points to different combinations of genes increasing susceptibility to schizophrenia. Subtle differences in the structure of the brains of some people with schizophrenia have been detected, although they did not imply causality [59].

Schizophrenia onset is usually characterized by a prodromal stage, with nonspecific symptoms stretching over several years, often creating social lifelong consequences. For male patients, the diagnosis of schizophrenia is more common in adolescence to mid-twenties, and for females, it extends to the early thirties with a lower peak at menopausal age [22]. Nevertheless, schizophrenia may manifest itself at all ages, and both sexes appear to be at equal lifelong risk.

Around 24 million people suffer from schizophrenia globally. In other words, for every 300 people, one person is schizophrenic. People with the disorder are subjected to severe and pervasive stigma, which makes them socially excluded, with negative consequences on how they connect with others. The prejudice against schizophrenia can restrict access to housing, education, and employment. Human rights abuses frequently occur to people with schizophrenia, both inside mental health facilities and in public places [38].

Throughout the world, people with schizophrenia have a higher mortality rate than the general population. Schizophrenia is a condition with a high prevalence of co-occurring illnesses, including diabetes and heart disease. People with schizophrenia have a 10% rate of suicide which is correlated with social context and cultural stigma around the disorder [46]. A way people with schizophrenia have found to deal with their challenges is to form online communities for peer support [12]. Nowadays, a variety of mental health communities are available on social media platforms, where people can express themselves anonymously.

2.2 Characteristics of Language in Schizophrenia

One predominant characteristic of schizophrenia is the impaired ability to maintain coherent communication, either verbally or in writing. This is related to difficulties in processing and organizing thoughts. This might be perceptible in the content of patients' speech, which might display delusions and hallucinations. Also, their discourse is disorganized, presenting lexical and syntactic abnormalities. Such cognitive, linguistic, and communicative oddities are referred to as formal thought disorder in the 11th edition of the International Classification of Diseases, organized by WHO - World Health Organization. The 5th edition of the DSM - Diagnostic and Statistical Manual, organized by the American Psychiatric Association, makes no reference to formal thought disorder, rather using the expression disorganized speech [40]. This reveals that there is a tight connection between formal thought disorder and language impairment in schizophrenia. Nevertheless, other cognitive resources, such as executive functions, including attention and memory, seem to be affected as well.

Linguistic inconsistencies produced by people with schizophrenia can be either a negative symptom when it is related to alogia, a diminishment of verbal production, or a positive symptom with disorganized or delirious discourse. As a positive symptom, both language production and comprehension might be impaired at a structural level, affecting different grammatical levels, including semantics, pragmatics, syntax, and morphology [28]. The presence of either positive or negative linguistic symptoms is not necessarily correlated with psychotic ideation. Whether accompanied by psychosis or not, language in schizophrenia often presents shorter statements, infrequent conjoined clauses, fewer words, and inconsistent verb forms, among other disturbances [13].

Schizophrenic discourse is also marked by fewer cohesive links of reference conjunction and lexical cohesion [31]. In fact, the analysis of language connectedness contrasting the disconnected content from people with schizophrenia to the loosely connected discourse of people with mania has proven to serve as a differential diagnosis of the conditions [36]. As for semantic fluency, people with schizophrenia are able to resort to fewer words indicating working memory impairments [4]. People with schizophrenia also display difficulty understanding metaphor, irony, idiomatic expressions, and other forms of figurative language [51].

Some linguistic symptoms have been related to specific cognition disruptions. People with schizophrenia often fail to clarify to whom they are referring when using pronouns, sometimes even mistakenly referring back to themselves using a third Person or second-person pronoun, a fact that has been connected to a lack of social cognitive abilities, particularly the theory of mind since to correctly use a pronoun that provides the listener with needed information, a person must hold a representation of the listener's mind [7].

Moreover, referentiality anomalies among schizophrenics are significantly higher within pronominal expressions, affecting 3rd person pronouns more than 1st and 2nd person pronouns and also impacting covert pronouns more than overt pronouns [52]. The preference for the use of interpersonal pronouns instead of proper nouns, and the overuse of 1st person pronouns, are also prevalent in linguistic expressions of people with schizophrenia [25]. A particularly increased use of 1st and 2nd person pronouns at the patient-specific level was correlated to a higher risk of hospitalization relapse [5].

A well-accepted theory in the research of schizophrenia is that language is more than a symptom. If we establish the premise that language and cognition are fundamentally integrated, we can approach grammar as an essential framework for a person's experience in the world. In that sense, a person's understanding and communication are mediated by language to the point that lexical components can be directly connected to specific cognitive functions. Thus if language disintegrates, so does the mental ability to process experience coherently. Hence, delusions, hallucinations, and other core symptoms of schizophrenia disorder would be a direct consequence of the fragmentation of a brain's language faculty [24].

2.3 Language as a Biomarker

The World Health Organization (WHO), the United Nations, and the International Labor Organization, in the International Programme on Chemical Safety, have stated that a biomarker is *"any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease."* [37]. Since clinicians mostly rely on language to diagnose and treat mental disorders, language impairments are considered to be an important potential biomarker.

We should point out that, as academic research using NLP methods has amplified the inherent value of linguistic biomarkers, some objections about the employment of the term biomarker in this particular context have emerged. Due to the undeniable social component of language, the term biosocial marker has been suggested considering that social factors affect linguistic markers significantly more than they affect the other markers of psychosis [39]. We do not object to the term biosocial marker or biomarker as referring to language measurements. Nevertheless, we decided to use the term biomarker based on the fact its use has been more widespread.

The advantages of using language as a biomarker are many, starting with the fact that NLP applied to the research of schizophrenia biomarkers has highlighted patterns that are indicative of particular diagnoses and symptoms, predictive for future outcomes, in addition to being indicators of disorder development and treatment outcomes. Previous experiments have shown that language analysis can account for either positive or negative symptoms and be effective in different stages of the disorder.

Moreover, the measurement of language is noninvasive in nature and potentially low-cost and time efficient. Language production can be objectively and reproducibly quantified, both in and out of clinical contexts, and it can even be further fitted to individual circumstances such as social context, medical history, and previously produced linguistic data.

2.4 NLP and Schizophrenia Research

Natural Language Processing techniques have been applied to the understanding of different aspects of both written and spoken language disorders in schizophrenia. In previous studies, NLP was proven to be a powerful resource to examine and describe formal thought disorder in terms of semantic coherence, connectedness, and other communication aspects.

Latent Semantic Analysis (LSA) has been explored as a method for measuring coherence. LSA is a statistical method for calculating the similarity between texts based on neighboring words. A study conducted by Elvevåg et al. [15] used LSA to represent pairs of questions and answers and compare them by computing cosine similarities between vector representations. The cosine similarities linear regression line slope would be a measure of response coherence. Therefore, the pronounced line rises at a very sharp angle, indicating that the answer was distancing from the question, suggested an increase in incoherence at the discourse level.

Another LSA approach was used by Bedi et al. [2] to predict the onset of psychosis. They calculated similarities between pairs of sentences separated by an intervening sentence to derive a global coherence estimation. Their work also involved extended features extracted with POS-tag and other syntactic features. They concluded that the minimum coherence between two consecutive phrases, maximum phrase length, and use of determiners was significantly correlated with prodromal symptoms. Nevertheless, one noticeable limitation of LSA methods for coherence detection is that repetition, a possible characteristic of formal thought disorder, would be registered as a great mark of coherence.

An analytic application tool used for describing formal thought disorder in schizophrenia is Linguistic Inquiry and Word Count (LIWC). LIWC applies a token-based method containing several dictionaries each reflecting a psychological state such as sadness, positive affect, and sociability. It counts how many words from each category occur in a given text. Buck et al. [8] found, using analysis of LIWC features, that even though people with schizophrenia and controls used a comparable number of words overall, the amount of words per sentence was a good predictor of schizophrenia. Finenberg et al. [16] calculated correlations between LIWC's cognitive and perceptual categories and used bivariate analysis with Pearson's correlation coefficient to assess first-person accounts from the journal Schizophrenia Bulletin. They uncovered that the correlation between causal and perceptual words is abnormal in reports of long-held delusions, indicating less sensory experience and causality awareness connection.

In another experiment, Mitchell et al. [33] compared 72 LIWC categories scores between control and schizophrenia groups and found that people with schizophrenia produced significantly more words from the cognitive mechanisms, death, function words, and negative emotion categories and fewer words from the home, leisure, and positive emotion categories. Additionally, they applied Latent Dirichlet Allocation (LDA), which can be considered a Bayesian approach to LSA assumptions, and discovered that the relative frequencies of the topics were considerably distinct in each group. In recent research, Zomick et al. [63] extracted LIWC features from Reddit posts to uncover that people with schizophrenia used more words related to health issues, anxiety, negative emotions, and first-person singular pronouns than controls.

Mota et al. [36], employed speech graphs to measure structural speech differences among people with mania and schizophrenia. The speech graphs represented language as a network with nodes corresponding to words and directed edges representing semantic and grammatical relationships. They concluded that quantitative analysis of speech graphs was able to sort people with mania from people with schizophrenia. They noted that when considering other developed psychometric scales, their results were not redundant, for they were able to measure speech structure symptoms that were not previously satisfactorily measured.

In our experiment, we approached language as a structural system using part-of-speech features to represent the text. We used two Machine Learning algorithms to classify the texts as being from the Schizophrenia group (henceforth SZ) or the control group (henceforth CT), and applied explainability techniques to explore the results. Additionally, we trained the models with three different hierarchical levels of word function description with distinct abilities to extract global or localized specifications, and combined the insights from each model to identify previously described SZ characteristics and propose new hypotheses. This approach allowed for nonlinear relations among features to emerge, adding another point of view to the description of POS features through frequency and correlation.

Chapter 3

Machine Learning Background

3.1 Machine Learning methods

Machine Learning ((henceforth ML) refers to the usage of algorithms and statistical models that computes data patterns to improve in a certain task without receiving any explicit instruction [34]. Not only the application of ML is useful in tasks that are difficult to be directly programmed, as it can also uncover patterns not previously known.

When using learning algorithms, we do not specify detailed steps required to achieve the desired outcomes. Instead, we provide our algorithm with a large number of examples so it can learn patterns related to the task. In terms of structure and logic, a learning algorithm's model is very different from a manually constructed rule-based system. The model learning process can be either supervised, unsupervised or semi-supervised [58].

Essentially supervised learning is done with a label that indicates the correct answer, or the ground truth, and by comparing its prediction to the label, the model can adjust through confirmation or contradiction. Supervised learning includes every task that requires access to an input and output value. On the other hand, unsupervised learning is the process of teaching the machine to respond to unlabeled data being restricted to identify the hidden structure of the data by itself. Without any prior data training, the machine's objective in this case is to categorize the imputed unsorted data according to similarities, patterns, and differences. At last, semi-supervised machine learning is a combination of supervised and unsupervised methods, a portion of the training data is matched with the correct output, but the rest is unlabeled. Basically, data is treated differently based on whether it is labeled or unlabeled. When dealing with labeled data, the algorithm uses a supervision approach and updates the model, whereas, for unlabeled data, the algorithm minimizes predicted differences among similar data. In the present study, we use supervised learning.

To ensure that a model is successful, it is important to establish whether a good result is due to the model's generalization ability or mere memorization of data. For that, the data must be split between the training dataset and the testing dataset before modeling. Training data refers to the information used as the model's first set of instances. Once trained, this model ought to be able to generalize and perform just as well when provided with new cases from the test dataset.

The outline of a ML experiment includes 6 main steps: data collection and preparation, feature engineering and selection, choosing the algorithm and training, evaluating the model, tuning the model, and explainability [44]. The first step refers to the acquisition, cleaning, and treatment of the data, being followed by data analysis and labeling and necessary data manipulation for the creation of features and the election of the most relevant features. The choice of the algorithm is typically directed by the problem requirements and highly influenced by data complexity and size, sometimes it is also limited by computational resources. Model tuning refers to the optimization of hyperparameters and other adjustments to get the best results, and is both preceded and followed by training and evaluation of results. A project might iterate through these 3 steps many times until it reaches a satisfactory result. Explainability is not always included in ML projects, although it helps characterize model outcomes, accuracy, fairness, transparency, and other desired model properties. It might also give insight into necessary adjustments [42].

3.2 The Bias-Variance Tradeoff

Two key concepts in ML are bias and variance. Variance relates to a model's incompetence to generalize what was learned from the training data and apply to new data with accuracy. Bias, on the other hand, tells us how well a model fits the training data, being that the better the fit of a model to the data, the lower its bias. When adjusting a model, we want to minimize both bias and variance. In other words, we want a model that matches the training data very well and gets the same result when testing new data. However, minimizing both properties simultaneously is not feasible because they are inversely related [41].

When the model presents minimal bias, being so well fitted to the training data that it almost memorized the training input, giving promising results in training but failing to apply the learning to new data due to high variance, we call that overfitting. The opposite situation, underfitting, would be to create a model that has minimal variance and is generalizing well but its training results did not live up to the potential due to high bias. Besides the bias and the variance error, a model can have irreducible data errors that cannot be predicted, which is referred to as noise. This type of error cannot be improved or fully removed, as it is inherent to the problem or caused by statistical noise in the data.

The bias and variance can vary with the algorithm chosen to train the model, the hyperparameters, the size of the data, and the features. An approach to deal with this trade-off is observing the difference between training and test results, and how much bias can be improved before variance gets too high. To estimate in the training phase the skill of a machine learning model on unseen data, we can use a cross-validation technique, which leads to a less biased or even a less optimistic estimate of the model capacity. In essence, cross-validation enables us to train a model using different samples from the training dataset. When we cross-validate, every input from the dataset appears once in a validation set. It's comparable to performing an experiment several times on several distinct datasets, developing a new model each time, and then averaging the outcomes. The cross-validation score is a reliable metric for understanding what can be expected from the model's performance on unseen data [3].

In our work we considered that texts from people with SZ can have both regular language characteristics and traits specific to SZ. It might even have more regular language characteristics at some point since SZ language anomalies follows the severity of the disorder, and that may vary over time. In other words, the data has noise that is inherent to the problem. For this reason we do not expect results to be significantly above the ones reported by previous works. Nonetheless, we aim for models with as low variance as possible, without increasing notably training bias.

3.3 Tree-Based Models

Tree-based classification algorithms are built on the fundamental idea that by learning a series of questions that divide situations repeatedly into specific subgroups they can achieve a good generalization of the classification problem. All questions have a binary response, and depending on which requirements are met, an input is routed along the left or right branch, creating branches within branches. A single decision tree can be easily interpreted by its graphical representation [11]. Its structure begins with a root node and then includes sequences of nodes where the split of the data occurs, branches leading from a node to another, and leaves that name the final nodes, which do not divide any further.

When a tree algorithm is trained for each node, the available features are considered one by one, and the one with the best discriminative power is chosen. A risk with decision trees is to create an over-complex model that does not generalize well from the training data. Another disadvantage is the fact that, since it is a greedy algorithm, it is not granted to reach a globally optimal decision tree [55].

3.4 Deep Learning

Deep learning refers to machine learning techniques that employ a multi-layered artificial neural network architecture. The principle of neural networks is to mirror the behavior of the human brain, using to this effect layers of interconnected nodes that are a simplified version of neurons [29]. Each layer can learn a different property of the data and sequentially they can deal with complex data patterns.

Even though artificial neural networks are much simpler than a human brain, they still perform tasks such as classification with powerful precision. The input moves through the hidden hierarchical layers in a non-linear order analogous to the human decision-making process and builds an understanding of the data from low-complex features to abstract concepts. In the learning processes, depending on the influence of a node on other nodes, it receives a particular weight that can be adjusted depending on the difference between the predicted output and its label [19].

The traditional feed-forward neural network usually fails to keep track of sequential relationships in data, which is an important property of textual data. Recurrent neural networks (RNN) offered a solution to this Natural Language Processing problem. For any imputed sequence of words, an RNN processes the first word and inputs the result into a layer that processes the next word. This allows the model to keep track of the entire sentence instead of processing words separately. Yet they can not handle long sentences, and the longer a sentence the bigger the risk of having an effect of the-first-word failure, a problem that is known as vanishing gradients. Another important limitation is that the RNNs only captures relationships of words that were immediately close to each other [61]. The successor of RNNs were Long short-term memory (LSTM), which solved to some degree the vanishing gradient problem, with improvements in advanced model variations as the bi-LSTM [62].

The Transformers architecture became the go-to solution since its introduction in the paper Attention Is All You Need [54]. They apply an attention mechanism that allows tracking-word relationships even across long sequences of text. We describe Transformers further in the next subsection.

3.4.1 Transformers

In this study, we use a transformer model to establish the discriminative power baseline for the other models. Since transformers are the state of the art in NLP projects [18], it serves as a reliable reference for the present limit of the data and task classification potential. This will be an important parameter to later evaluate the explainability and accuracy trade-off from our proposed modeling approach.

An important addition to the transformer architecture was the employment of attention layers vaswani2017attention. When the word embeddings are imputed into the transformers encoder, they are processed in parallel instead of sequentially, and the sequential property of the input is handled with the use of positional encodings that represents the location of the embedding in the text.

An attention layer is applied to try and capture the associations of embeddings in the context of the input. It receives the word embeddings and produces new embeddings that reflect both the words and the relations among them. Then a neural network processes the attention layer output embeddings and feeds the results to another attention layer. A transformer might have several blocks of attention and feed-forward layers, and this is how it is able to apprehend highly complex relationship patterns in the data.

At the end of a transformer architecture, a decoder module receives the output of the last attention layer and translates it to the final model output. In the training phase, the decoder processes the expected outcome and, according to the difference of the label and the model's output, it creates attention vectors. It then passes this output to the encoder module, which establishes relations between the input and output values.

3.5 Model Explainability

The present investigation uses the terms explainability and interpretability, interchangeability to refer to the degree to which a model output can be demonstrated and understood by humans. This is a way of guaranteeing low bias, providing a way to explore the results in search of new patterns or insights on the problem posed while indicating that the model's reasoning is trusting worth

Machine Learning models can be directly interpretable, but they might also be limited to *post hoc* explanations [35]. The second possibility refers to the process of probing the model to derive the conclusion of which parts or aspects of the input were most relevant to the definition of the output. Explanation methods can be only applicable to a model or a group of models, or they can be model-agnostic.

When a technique is used to explain the model as a whole, it is considered a global interpretation, when seeking to define explanations only for a particular input it is called a local explanation. Sometimes achieving global explainability is difficult, considering that a model's reasoning might involve a fair amount of complex patterns that are conditionally applied to the data depending on other data aspects and humans cannot easily hold that much information in memory. A model with many features or with a complicated multifaceted problem can often be hard to explain.

3.5.1 Layer Integrated Gradients

Layer Integrated Gradients is an explainability method applicable to Transformers where the attribution scores are a summary or average of a layer's attributions. It is a *post-hoc* technique that outputs a score for each feature indicating how a feature contributed to the model's output [49]. That means that a positive score indicates that the feature agrees with the model's classification, while a negative score indicates that the feature disagrees with it.

The integrated Gradients method starts without any information on the true output and then it obtains the output for different sections of the input. A simplified way to explain it is to consider that it starts on the first word and recursively add the next one getting the layer output in each step, forming a list of attributions value for each feature. If an input feature alters the output in any way, this input should have an attribution value different from 0.

3.5.2 Shapley values

A Shapley value is calculated as the average marginal contribution of a feature value to the prediction across all possible associations of features [48]. A fundamental property of Shapley values is their additive nature since they always sum up to the difference between the output when all features are present and the output when no features are present. If two features present the same contribution in all combinations of features their calculated Shapley value is the same, a property called symmetry, and in the case that a feature does not change results no matter which combination of features being tested, its value is zero.

Another advantage of the method is that the calculated Shapley values admit contrastive explanations. It is fair to compare two local explanations or even local explanations to the explanations of different subsets of data. This contrastive attribute is not granted in all interpretability methods.

We should point out that global Shapley's values can easily be misinterpreted as the difference in the prediction score after a feature is removed from a model. An accurate way of thinking would be that it is the contribution of a feature value to the difference of the ground truth prediction and the one estimated by the method given the entire model's feature set [32].

A disadvantage of the process could be that if you want to calculate the Shapley value for a new input not used for training, it might not be enough to have access to the prediction function to get the prediction score, being necessary to have access to the training data. This is due to the fact that the simulation that a feature value is missing from a combination of features is achieved by sampling values from the feature's marginal distribution. So it would be necessary to access the data to sample the replacement for the parts of the input.

Chapter 4

Data and Algorithms

4.1 Overview of Research Method

This study aims at contributing to the research of language as a cognitive biomarker for schizophrenia. It was carried out by the development of machine learning classification models using sets of lexical features with different levels of specificity. We collected the data and created a specialized corpus for the task. We used a POS tagging model and developed a function to further suit the tags to our proposition. We engineered 3 levels of features and selected 4 sets of features with a supervised approach. As a baseline for classification performance, a state-of-the-art transformer model was trained in the same task. Using two different machine learning algorithms and the selected features of the 4 sets, 8 models were trained and tested. We analyzed the results considering model robustness and discriminative power, and also the model explainability potential for contribution to psycholinguistic research. We detail in this chapter the experiment steps and strategies mentioned. To facilitate comprehension of the procedure, we outlined the experiment workflow in figure 4.1

4.2 Corpus Creation and Preparation

4.2.1 Data Source

The data was acquired through a web scraping process of selected content on the Reddit platform. Reddit is a social news website and forum where content is socially curated and promoted by the site's pseudonymous members through voting. Reddit is one of the most current relevant social media, being recorded in March 2022 as the 9th-

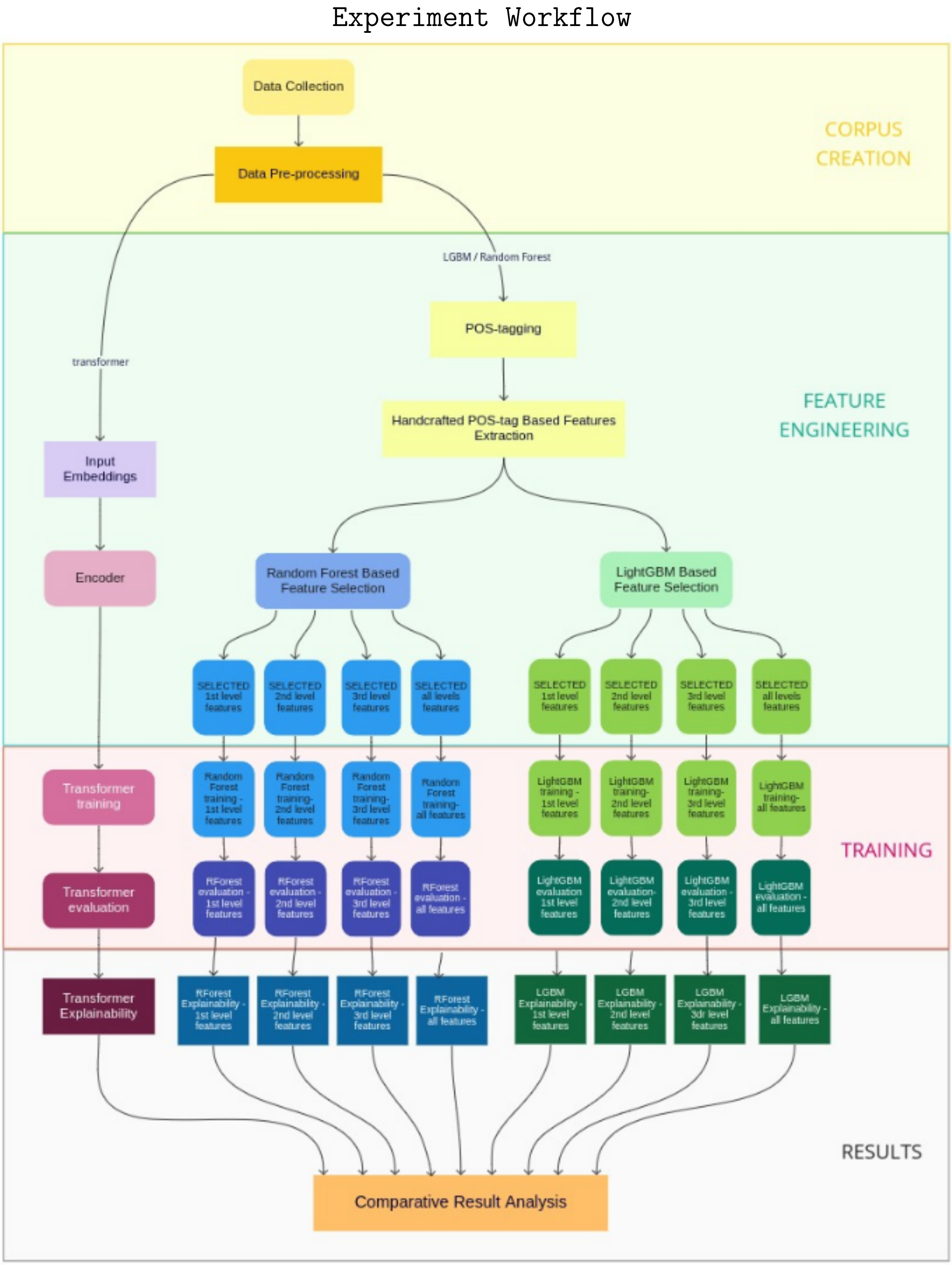

Experiment Workflow
Data Collection
Data Pre-processing
CORPUS CREATION
LGBM / Random Forest
transformer
POS-tagging
Handcrafted POS-tag Based Features Extraction
FEATURE ENGINEERING
Input Embeddings
Encoder
Random Forest Based Feature Selection
LightGBM Based Feature Selection
SELECTED 1st level features
SELECTED 2nd level features
SELECTED 3rd level features
SELECTED all levels features
SELECTED 1st level features
SELECTED 2nd level features
SELECTED 3rd level features
SELECTED all levels features
Transformer training
Random Forest training - 1st level features
Random Forest training- 2nd level features
Random Forest training- 3rd level features
Random Forest training- all features
LightGBM training - 1st level features
LightGBM training- 2nd level features
LightGBM training- 3rd level features
LightGBM training- all features
TRAINING
Transformer evaluation
RForest evaluation - 1st level features
RForest evaluation - 2nd level features
RForest evaluation - 3rd level features
RForest evaluation - all features
LightGBM evaluation 1st level features
LightGBM evaluation- 2nd level features
LightGBM evaluation - 3rd level features
LightGBM evaluation - all features
Transformer Explainability
RForest Explainability - 1st level features
RForest Explainability - 2nd level features
RForest Explainability - 3rd level features
RForest Explainability - all features
LGBM Explainability - 1st level features
LGBM Explainability - 2nd level features
LGBM Explainability - 3dr level features
LGBM Explainability - all features
RESULTS
Comparative Result Analysis

Figure 4.1: Outline of steps of the experiment workflow

most-visited website in the world and the 6th most-visited website in the United States. It is fair to assume that most of Reddit users speak English as a first language since up to 49.3% of Reddit's user base is located in the United States, followed by the United

Kingdom at 7.9 to 8.2% and Canada at 5.2 to 7.8%, according to Semrush [45].

In addition to its relevance, we chose to work with Reddit data for its forum characteristics, which gave us a chance to flag self-identified schizophrenic users to set a target group. Reddit forums are user-created areas called Subreddits where discussions of a specific topic are organized. Reddit users are called Redditors, they can enter different Subreddits of their choice, and they are free to post new topic-related content or comment on other user's posts to continue a conversation. Redditors also can cast positive or negative votes for each post or comment, and these votes determine visibility on the site, meaning that the most popular content is displayed to more people. Each subreddit might have a unique tagging system consisting of limited sets of tags called flairs that can be associated either to a user or a post.

The use of social media data in scientific research has both advantages and disadvantages. Since it is not as carefully collected as it would be in a more conservative controlled experiment, many aspects cannot be controlled for, and some of the conclusions might be limited to further testing and considerations. On the other hand, social media data is widely available, not only providing the data in the amount necessary to satisfactorily train ML models but also reducing the experiment's cost and time. In previous NLP works, the use of Reddit data has shown to be both practical and insightful [6]

4.2.2 Data Collection

For web scraping, we used an API provided by Reddit named pushshift.io [1], which provides enhanced capabilities for searching Reddit comments and submissions. There's a limit of 100 posts retrieved by month and old posts without any upvoting might be archived by Reddit and not appear in any type of search. We started at the Schizophrenia Subreddit and attempted to collect posts from the period starting the beginning of 2009 until the 10th of February of 2021. We were not able to get any posts from before 2012.

In the Schizophrenia subreddit, one of the available user tags is the schizophrenic flair, and it was through this flair that we flagged self-identified schizophrenic Redditors. We selected those Redditors and visited their profile to access all their posting history. We then collected all available selected user's posts, including the comments, on all the subreddits they took part in, from the period of time starting at the beginning of 2017 until march 2021. This concluded the collection of the target group data (SZ).

For the collection of the control group data, since we could not control if there were people with schizophrenia among redditors of the CT group, to prevent bias, we gave preference for collecting posts from a wider range of redditors instead of trying to reproduce the proportion of posts per redittors of the target group. The criterion used was subreddit representation, and ten subreddits were selected considering the distribution of the subreddits in the target group that was most representative in topic and size. From these subreddits, we searched posts from redditors that were not in the SZ group.

Since the CT group date was easily available, it was not necessary to expand the time range of the search before January 2019, and the latest searched date was also March 2021.

4.2.3 Subreddits

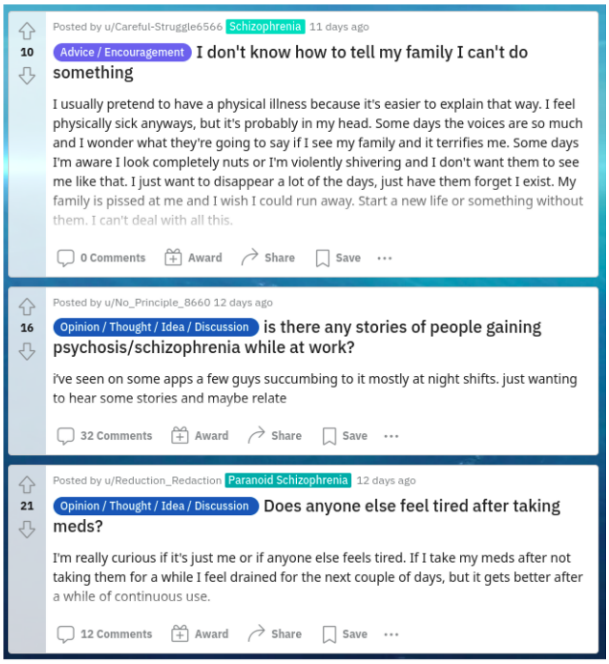
Being subreddits a important criteria in our data collection methodology , and given the fact that posts from distinct subreddits have different themes and might even vary in pattern, it is relevant to mention and briefly describe the subreddits from which most posts represented in our sample were collected. They are:

Schizophrenia:

- This is a subreddit for the schizophrenia community including people diagnosed with schizophrenia, people that are close to people living with schizophrenia, and health professionals. Content includes questions, testimonials, encouragement, requests for support, humor, art, selfies, medical facts, and other schizophrenia-related content. Figure 4.2 shows an excerpt from the *Schizophrenia* subreddit main page.
- Self-description: *Welcome! This is a community meant for a discussion of schizophrenia spectrum disorders and related issues. Active participation is encouraged.*
- Page URL: https://www.reddit.com/r/schizophrenia/
- Total of members: 58.6k

Ask Reddit:

- This subreddit consists of questions and answers on a variety of topics, most of them controversial or stimulating. Figure 4.3 shows an excerpt from *Ask Reddit* main page.
- Self-description: *AskReddit is the place to ask and answer thought-provoking questions.*
- Page URL: https://www.reddit.com/r/AskReddit/



Posted by u/Careful-Struggle6566 Schizophrenia 11 days ago

10

Advice / Encouragement I don't know how to tell my family I can't do something

I usually pretend to have a physical illness because it's easier to explain that way. I feel physically sick anyways, but it's probably in my head. Some days the voices are so much and I wonder what they're going to say if I see my family and it terrifies me. Some days I'm aware I look completely nuts or I'm violently shivering and I don't want them to see me like that. I just want to disappear a lot of the days, just have them forget I exist. My family is pissed at me and I wish I could run away. Start a new life or something without them. I can't deal with all this.

0 Comments Award Share Save ...

Posted by u/No_Principle_8660 12 days ago

16

Opinion / Thought / Idea / Discussion is there any stories of people gaining psychosis/schizophrenia while at work?

i've seen on some apps a few guys succumbing to it mostly at night shifts. just wanting to hear some stories and maybe relate

32 Comments Award Share Save ...

Posted by u/Reduction_Redaction Paranoid Schizophrenia 12 days ago

21

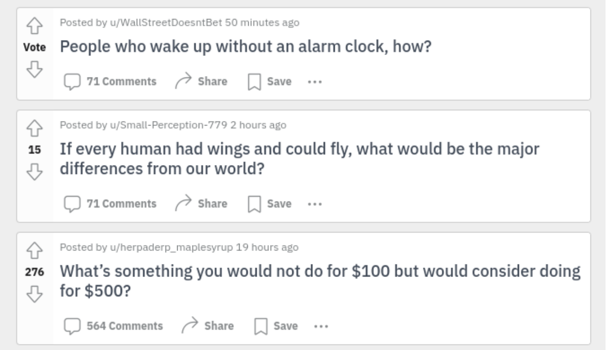
Opinion / Thought / Idea / Discussion Does anyone else feel tired after taking meds?

I'm really curious if it's just me or if anyone else feels tired. If I take my meds after not taking them for a while I feel drained for the next couple of days, but it gets better after a while of continuous use.

12 Comments Award Share Save ...

Figure 4.2: Excerpt from the *Schizophrenia* subreddit feed

- Total of members: 37.1m



Posted by u/WallStreetDoesntBet 50 minutes ago

Vote

People who wake up without an alarm clock, how?

71 Comments Share Save ...

Posted by u/Small-Perception-779 2 hours ago

15

If every human had wings and could fly, what would be the major differences from our world?

71 Comments Share Save ...

Posted by u/herpaderp_maplesyrup 19 hours ago

276

What's something you would not do for $100 but would consider doing for $500?

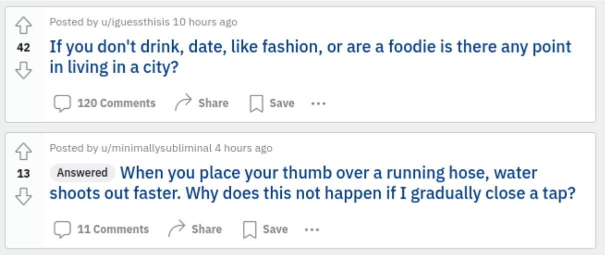
564 Comments Share Save ...

Figure 4.3: Excerpt from the *Ask Reddit* subreddit feed

No Stupid Questions:

- This is a question-and-answer subreddit with provocative, informative, and fact-checking topics. Figure 4.4 shows an excerpt from *No Stupid Questions* main page.
- Self-description: *Ask away! No such thing as stupid questions*
- Page URL: https://www.reddit.com/r/NoStupidQuestions/
- Total of members: 2.9m

Pokemon Sword And Shield:



Posted by u/iguessthisis 10 hours ago

42

If you don't drink, date, like fashion, or are a foodie is there any point in living in a city?

120 Comments Share Save ...

Posted by u/minimallysubliminal 4 hours ago

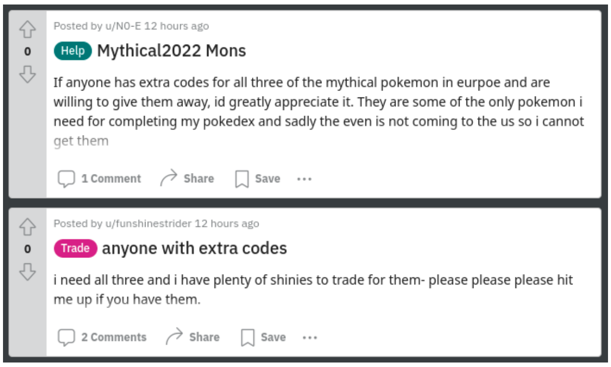
13

Answered When you place your thumb over a running hose, water shoots out faster. Why does this not happen if I gradually close a tap?

11 Comments Share Save ...

Figure 4.4: Excerpt from the *No Stupid Questions* subreddit feed

- A subreddit for players of the game *Pokemon Sword & Shield*, with posts consisting mainly of discussions of game features and strategies, requests for help, and trade. Figure 4.5 shows an excerpt from the subreddit's main page.
- Self-description: *A subreddit to discuss anything about Pokemon Sword & Shield!*
- Page URL: https://www.reddit.com/r/PokemonSwordAndShield/
- Total of members: 648k



Posted by u/N0-E 12 hours ago

0

Help Mythical2022 Mons

If anyone has extra codes for all three of the mythical pokemon in eurpoe and are willing to give them away, id greatly appreciate it. They are some of the only pokemon i need for completing my pokedex and sadly the even is not coming to the us so i cannot get them

1 Comment Share Save ...

Posted by u/funshinestrider 12 hours ago

0

Trade anyone with extra codes

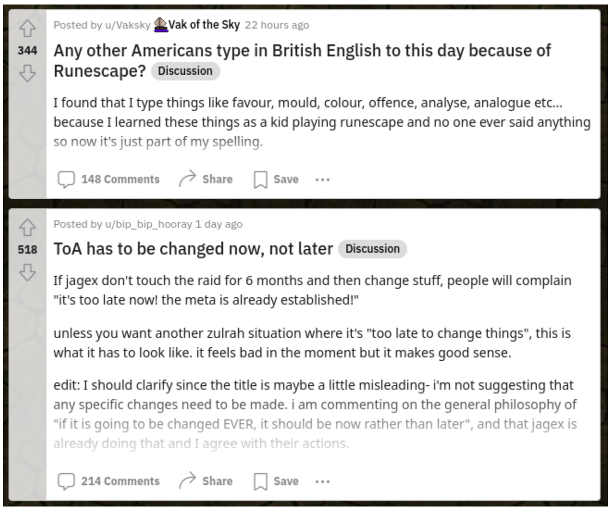
i need all three and i have plenty of shinies to trade for them- please please please hit me up if you have them.

2 Comments Share Save ...

Figure 4.5: Excerpt from the *Pokemon Sword And Shield* subreddit feed

2007 Scape:

- The subreddit consists mainly of achievement celebration, discussion, suggestions and humor related to the game *RuneScape*. Figure 4.6 shows an excerpt from the subreddit's main page.
- Self-description: *The community for Old School RuneScape discussion on Reddit. Join us for game discussions, tips and tricks, and all things OSRS! OSRS is the official legacy version of RuneScape, the largest free-to-play MMORPG.*
- Page URL: https://www.reddit.com/r/2007scape/
- Total of members: 670k



Posted by u/Vaksky Vak of the Sky 22 hours ago

344

Any other Americans type in British English to this day because of Runescape? Discussion

I found that I type things like favour, mould, colour, offence, analyse, analogue etc... because I learned these things as a kid playing runescape and no one ever said anything so now it's just part of my spelling.

148 Comments Share Save ···

Posted by u/bip_bip_hooray 1 day ago

518

ToA has to be changed now, not later Discussion

If jagex don't touch the raid for 6 months and then change stuff, people will complain "it's too late now! the meta is already established!"

unless you want another zulrah situation where it's "too late to change things", this is what it has to look like. it feels bad in the moment but it makes good sense.

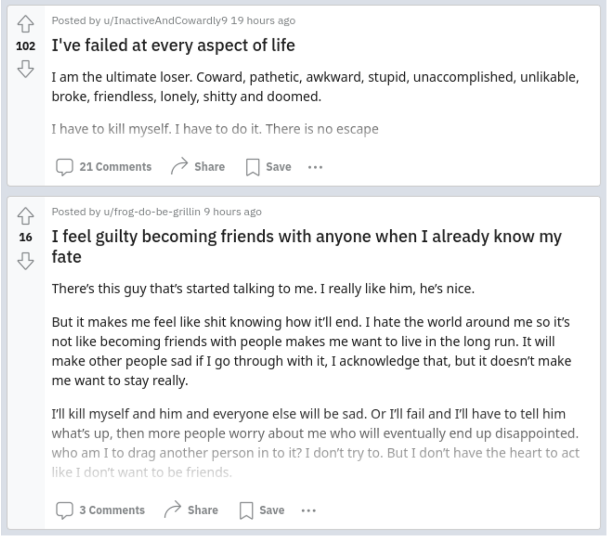
edit: I should clarify since the title is maybe a little misleading- i'm not suggesting that any specific changes need to be made. i am commenting on the general philosophy of "if it is going to be changed EVER, it should be now rather than later", and that jagex is already doing that and I agree with their actions.

214 Comments Share Save ···

Figure 4.6: Excerpt from the *2007scape* subreddit feed

Suicide Watch:

- In this subreddit, Redditors share suicidal thoughts and related mental sufferings. This subreddit is moderated and discussions around relatability, motivation or tips for improvement are allowed, while direct suicidal encouragement is forbidden. Moderators also discourage scientifical or technical posts even if they are suicidal-related. Figure 4.7 shows an excerpt from the subreddit's main page.
- Self-description: *Peer support for anyone struggling with suicidal thoughts.*
- Page URL: https://www.reddit.com/r/SuicideWatch/
- Total of members: 376k



Posted by u/InactiveAndCowardly9 19 hours ago

102

I've failed at every aspect of life

I am the ultimate loser. Coward, pathetic, awkward, stupid, unaccomplished, unlikable, broke, friendless, lonely, shitty and doomed.

I have to kill myself. I have to do it. There is no escape

21 Comments Share Save ···

Posted by u/frog-do-be-grillin 9 hours ago

16

I feel guilty becoming friends with anyone when I already know my fate

There's this guy that's started talking to me. I really like him, he's nice.

But it makes me feel like shit knowing how it'll end. I hate the world around me so it's not like becoming friends with people makes me want to live in the long run. It will make other people sad if I go through with it, I acknowledge that, but it doesn't make me want to stay really.

I'll kill myself and him and everyone else will be sad. Or I'll fail and I'll have to tell him what's up, then more people worry about me who will eventually end up disappointed. who am I to drag another person in to it? I don't try to. But I don't have the heart to act like I don't want to be friends.

3 Comments Share Save ···

Figure 4.7: Excerpt from the *Suicide Watch* subreddit feed

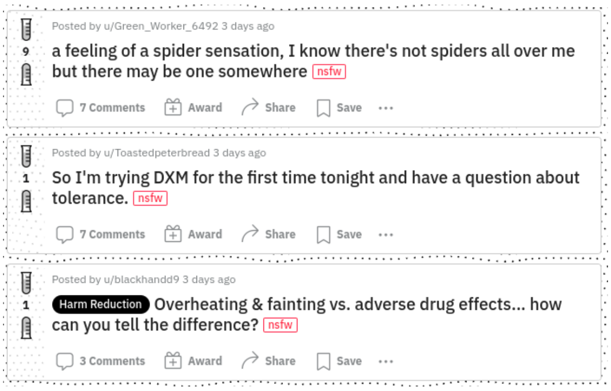
Drugs:

- This is a subreddit with questions, stories, and instructions related to drug use, including any type of drug, interactions, dosages, effects, or tips for enjoyable drug abuse with harm reduction. Posts might be controversial, technical, cautioning, or descriptions of self-experience. Figure 4.8 shows an excerpt from the subreddit's main page.

- Self-description: *We do NOT promote drug use. Accepts, for better and or worse, that licit & illicit drug use is part of our world and chooses to work to minimize its harmful effects rather than simply ignore or condemn them. Utilizing evidence-based, feasible, and cost-effective practices to prevent and reduce harm; Calls for the non-judgmental, non-coercive provision of services and resources to people who use drugs*

- Page URL: https://www.reddit.com/r/Drugs/

- Total of members: 902k



Posted by u/Green_Worker_6492 3 days ago

a feeling of a spider sensation, I know there's not spiders all over me but there may be one somewhere nsfw

Posted by u/Toastedpeterbread 3 days ago

So I'm trying DXM for the first time tonight and have a question about tolerance. nsfw

Posted by u/blackhandd9 3 days ago

Harm Reduction Overheating & fainting vs. adverse drug effects... how can you tell the difference? nsfw

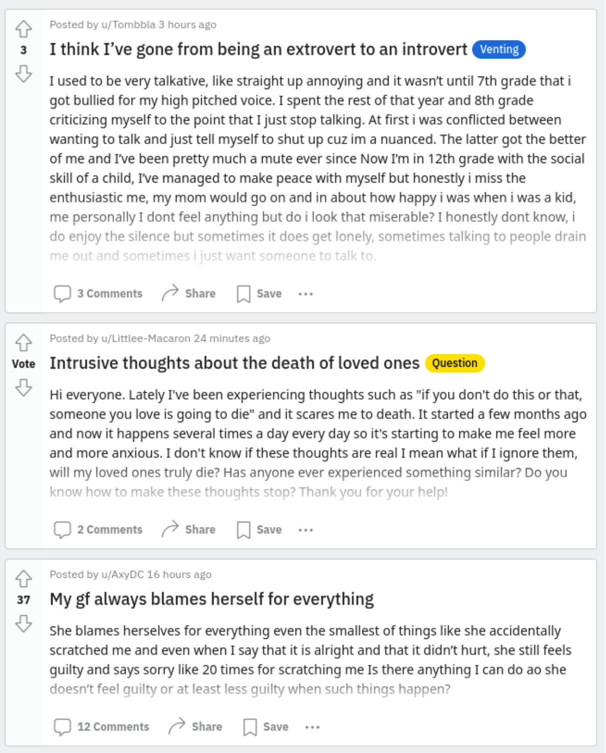
Figure 4.8: Excerpt from the *Drugs* subreddit feed

Mental Health:

- In this subreddit redittors can ask questions, vent about mental health struggles, ask for support, share good news, inspiration, sadness, or grief experience. Moderators prevent hate speech or another content considered harmful . Figure 4.9 shows an excerpt from the subreddit's main page.

- Self-description: *The Mental Health subreddit is the central forum to discuss, vent, support and share information about mental health, illness and wellness. This sub is moderated by the South Asian Mental Health Alliance (SAMHAA), a non-profit society dedicated to mental health stigma reduction through skill development and community building.*

- Page URL: https://www.reddit.com/r/mentalhealth/
- Total of members: 349k



Posted by u/Tombbla 3 hours ago

3

I think I've gone from being an extrovert to an introvert

I used to be very talkative, like straight up annoying and it wasn't until 7th grade that i got bullied for my high pitched voice. I spent the rest of that year and 8th grade criticizing myself to the point that I just stop talking. At first i was conflicted between wanting to talk and just tell myself to shut up cuz im a nuanced. The latter got the better of me and I've been pretty much a mute ever since Now I'm in 12th grade with the social skill of a child, I've managed to make peace with myself but honestly i miss the enthusiastic me, my mom would go on and in about how happy i was when i was a kid, me personally I dont feel anything but do i look that miserable? I honestly dont know, i do enjoy the silence but sometimes it does get lonely, sometimes talking to people drain me out and sometimes i just want someone to talk to.

Posted by u/Littlee-Macaron 24 minutes ago

Intrusive thoughts about the death of loved ones

Hi everyone. Lately I've been experiencing thoughts such as "if you don't do this or that, someone you love is going to die" and it scares me to death. It started a few months ago and now it happens several times a day every day so it's starting to make me feel more and more anxious. I don't know if these thoughts are real I mean what if I ignore them, will my loved ones truly die? Has anyone ever experienced something similar? Do you know how to make these thoughts stop? Thank you for your help!

Posted by u/AxyDC 16 hours ago

37

My gf always blames herself for everything

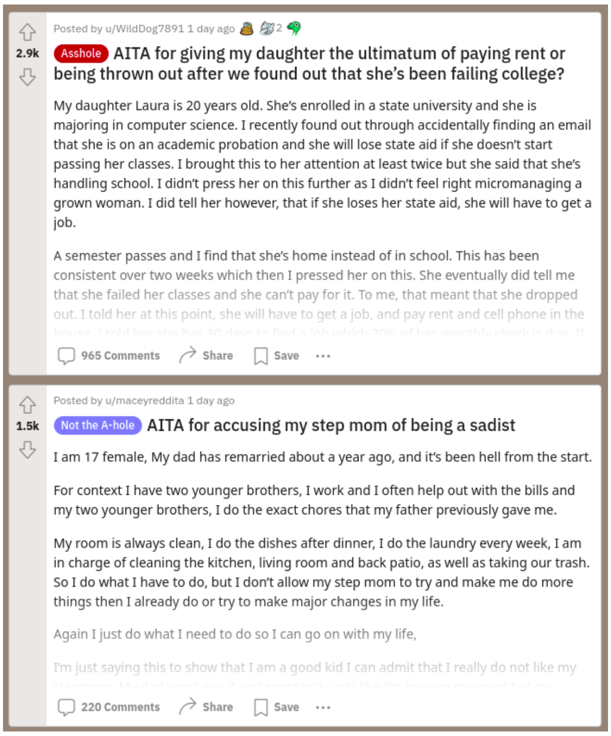
She blames herselves for everything even the smallest of things like she accidentally scratched me and even when I say that it is alright and that it didn't hurt, she still feels guilty and says sorry like 20 times for scratching me Is there anything I can do ao she doesn't feel guilty or at least less guilty when such things happen?

Figure 4.9: Excerpt from the *Mental Health* subreddit feed

Am I the Asshole:

- This is a subreddit where Redditors ask questions starting with AITA, which stands for "Am I the asshole?", and describe the context about the situation that prompted the question. The posts can be assessed by other redditors vote as: "not the a-hole", meaning the Redditor is considered to be right; "asshole", meaning the Redditor is considered to be wrong; "everyone sucks", meaning all involved in the situation are considered to be wrong; or even "no a-holes here", meaning everyone is considered to be right. Besides the vote, other Redditors can comment on the situation, give advice and share similar experiences. Figure 4.10 shows an excerpt from the subreddit's main page.
- Self-description: *A catharsis for the frustrated moral philosopher in all of us, and a place to finally find out if you were wrong in an argument that's been bothering you. Tell us about any non-violent conflict you have experienced; give us both sides of the story, and find out if you're right, or you're the asshole.*
- Page URL: https://www.reddit.com/r/AmItheAsshole/

- Total of members: 4.6m



Posted by u/WildDog7891 1 day ago 2

AITA for giving my daughter the ultimatum of paying rent or being thrown out after we found out that she's been failing college?

My daughter Laura is 20 years old. She's enrolled in a state university and she is majoring in computer science. I recently found out through accidentally finding an email that she is on an academic probation and she will lose state aid if she doesn't start passing her classes. I brought this to her attention at least twice but she said that she's handling school. I didn't press her on this further as I didn't feel right micromanaging a grown woman. I did tell her however, that if she loses her state aid, she will have to get a job.

A semester passes and I find that she's home instead of in school. This has been consistent over two weeks which then I pressed her on this. She eventually did tell me that she failed her classes and she can't pay for it. To me, that meant that she dropped out. I told her at this point, she will have to get a job, and pay rent and cell phone in the house. I told her she has 30 days to find a job which 30% of her monthly check is due. If

Posted by u/maceyreddita 1 day ago

AITA for accusing my step mom of being a sadist

I am 17 female, My dad has remarried about a year ago, and it's been hell from the start.

For context I have two younger brothers, I work and I often help out with the bills and my two younger brothers, I do the exact chores that my father previously gave me.

My room is always clean, I do the dishes after dinner, I do the laundry every week, I am in charge of cleaning the kitchen, living room and back patio, as well as taking our trash. So I do what I have to do, but I don't allow my step mom to try and make me do more things then I already do or try to make major changes in my life.

Again I just do what I need to do so I can go on with my life,

I'm just saying this to show that I am a good kid I can admit that I really do not like my stepmom. My dad won't see it and constantly acts like I'm loosing my mind but my

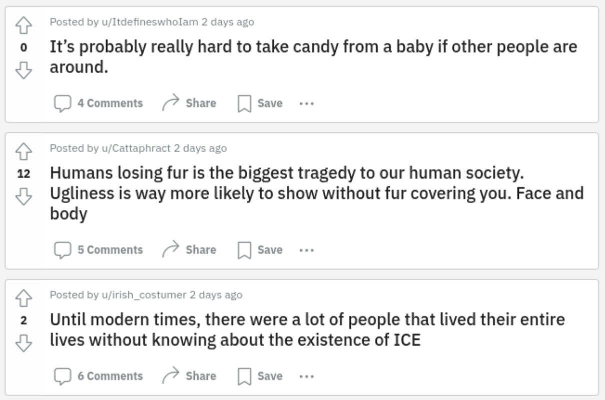
Figure 4.10: Excerpt from the *Am I the Asshole* subreddit feed

Shower Thoughtst:

- In this subreddit, redittors share insights about any topic they found curious, surprising, or tragic. Figure 4.11 shows an excerpt from the subreddit's main page.
- Self-description: *A subreddit for sharing those miniature epiphanies you have that highlight the oddities within the familiar.*
- Page URL: https://www.reddit.com/r/Showerthoughts/
- Total of members: 25.6m

Unpopular Opinion:

- In this subreddit posts always starts with a statement the redditor considers to be unpopular followed by the reasoning behind it. Other redditors can comment and expand on the topic. Figure 4.12 shows an excerpt from the subreddit's main page.
- Self-description: *Got a burning unpopular opinion you want to share? Spark some discussions!*
- Page URL: https://www.reddit.com/r/unpopularopinion/



Posted by u/ItdefineswhoIam 2 days ago

It’s probably really hard to take candy from a baby if other people are around.

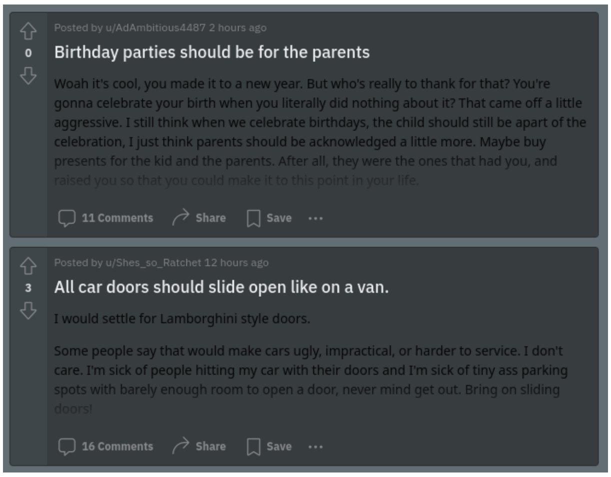
Posted by u/Cattaphract 2 days ago

Humans losing fur is the biggest tragedy to our human society. Ugliness is way more likely to show without fur covering you. Face and body

Posted by u/irish_costumer 2 days ago

Until modern times, there were a lot of people that lived their entire lives without knowing about the existence of ICE

Figure 4.11: Excerpt from the *Shower Thoughtst* subreddit feed



Posted by u/AdAmbitious4487 2 hours ago

Birthday parties should be for the parents

Woah it's cool, you made it to a new year. But who's really to thank for that? You're gonna celebrate your birth when you literally did nothing about it? That came off a little aggressive. I still think when we celebrate birthdays, the child should still be apart of the celebration, I just think parents should be acknowledged a little more. Maybe buy presents for the kid and the parents. After all, they were the ones that had you, and raised you so that you could make it to this point in your life.

Posted by u/Shes_so_Ratchet 12 hours ago

All car doors should slide open like on a van.

I would settle for Lamborghini style doors.

Some people say that would make cars ugly, impractical, or harder to service. I don't care. I'm sick of people hitting my car with their doors and I'm sick of tiny ass parking spots with barely enough room to open a door, never mind get out. Bring on sliding doors!

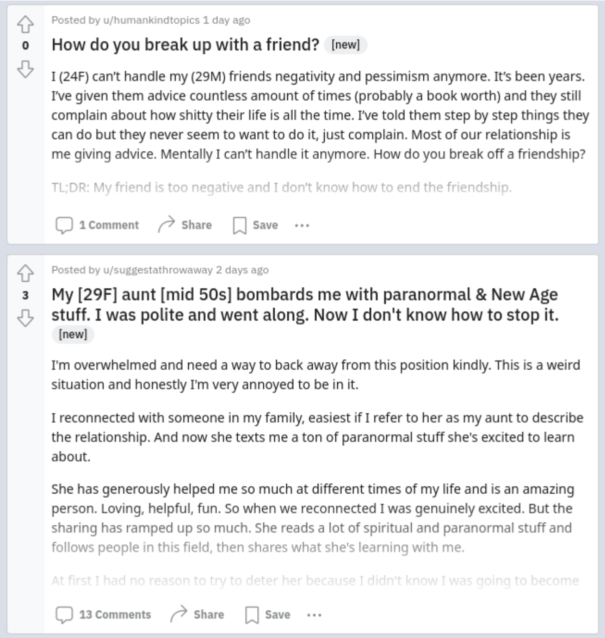
Figure 4.12: Excerpt from the *Unpopular Opinion* subreddit feed

- Total of members: 2.9m

Relationships:

- This is a subreddit where redditors share a problem related to a relationship with someone in order to receive opinions and advice. Figure 4.13 shows an excerpt from the subreddit’s main page.
- Self-description: *Relationships is a community built around helping people and the goal of providing a platform for interpersonal relationship advice between redditors. We seek posts from users with specific and personal relationship quandaries that other redditors can help them try to solve.*
- Page URL: https://www.reddit.com/r/relationships/
- Total of members: 3.3m

Neoliberal:



Posted by u/humankindtopics 1 day ago

How do you break up with a friend? [new]

I (24F) can't handle my (29M) friends negativity and pessimism anymore. It's been years. I've given them advice countless amount of times (probably a book worth) and they still complain about how shitty their life is all the time. I've told them step by step things they can do but they never seem to want to do it, just complain. Most of our relationship is me giving advice. Mentally I can't handle it anymore. How do you break off a friendship?

TL;DR: My friend is too negative and I don't know how to end the friendship.

Posted by u/suggestathrowaway 2 days ago

My [29F] aunt [mid 50s] bombards me with paranormal & New Age stuff. I was polite and went along. Now I don't know how to stop it. [new]

I'm overwhelmed and need a way to back away from this position kindly. This is a weird situation and honestly I'm very annoyed to be in it.

I reconnected with someone in my family, easiest if I refer to her as my aunt to describe the relationship. And now she texts me a ton of paranormal stuff she's excited to learn about.

She has generously helped me so much at different times of my life and is an amazing person. Loving, helpful, fun. So when we reconnected I was genuinely excited. But the sharing has ramped up so much. She reads a lot of spiritual and paranormal stuff and follows people in this field, then shares what she's learning with me.

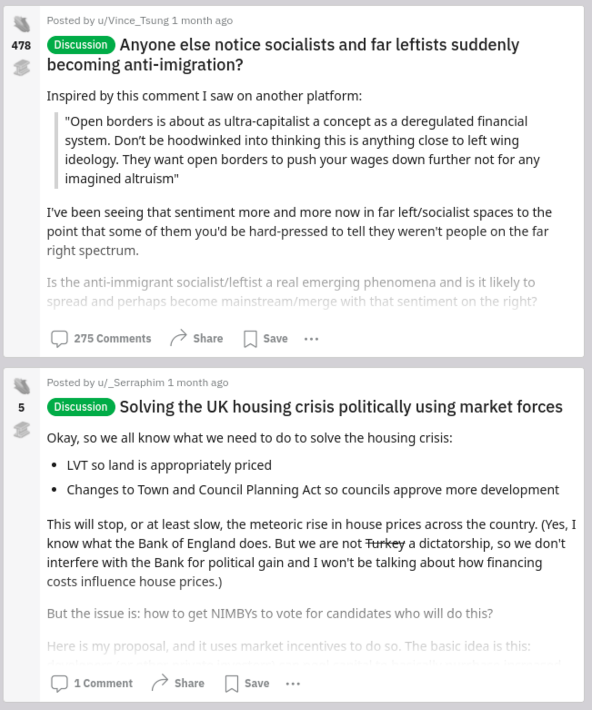
At first I had no reason to try to deter her because I didn't know I was going to become

Figure 4.13: Excerpt from the *Relationships* subreddit feed

- This is a subreddit with news and discussions pertinent to politics and the neoliberal point of view. Figure 4.14 shows an excerpt from the subreddit's main page.
- Self-description: *Free trade, open borders, taco trucks on every corner. Please read the sidebar for more information.*
- Page URL:https://www.reddit.com/r/neoliberal/
- Total of members: 138k

4.2.4 Data Preparation

In preparation for the POS-tagging and the encoding step, a few text pre-processing techniques were applied to standardize and normalize the text. This included removal of links, web addresses, extra spaces, and non-character elements such as emojis. We checked for duplicate posts both by post identification and identical text and excluded repeated occurrences. We did not remove the so called stop words (e.g; articles, light verbs, prepositions) or punctuation since removing them would not fit the purpose of our study, which is learning more about the data from a structural point of view. Given our interest on structure we also applied the criterion of 5-word minimum posts.



Posted by u/Vince_Tsung 1 month ago

478

Anyone else notice socialists and far leftists suddenly becoming anti-imigration?

Inspired by this comment I saw on another platform:

> "Open borders is about as ultra-capitalist a concept as a deregulated financial system. Don't be hoodwinked into thinking this is anything close to left wing ideology. They want open borders to push your wages down further not for any imagined altruism"

I've been seeing that sentiment more and more now in far left/socialist spaces to the point that some of them you'd be hard-pressed to tell they weren't people on the far right spectrum.

Is the anti-immigrant socialist/leftist a real emerging phenomena and is it likely to spread and perhaps become mainstream/merge with that sentiment on the right?

Posted by u/_Serraphim 1 month ago

5

Solving the UK housing crisis politically using market forces

Okay, so we all know what we need to do to solve the housing crisis:

- LVT so land is appropriately priced
- Changes to Town and Council Planning Act so councils approve more development

This will stop, or at least slow, the meteoric rise in house prices across the country. (Yes, I know what the Bank of England does. But we are not ~~Turkey~~ a dictatorship, so we don't interfere with the Bank for political gain and I won't be talking about how financing costs influence house prices.)

But the issue is: how to get NIMBYs to vote for candidates who will do this?

Here is my proposal, and it uses market incentives to do so. The basic idea is this:

Figure 4.14: Excerpt from the *Neoliberal* subreddit feed

4.2.5 Corpus Description

After collection and preparation, the final corpus was composed of 408,791 post's texts, being 151,320 in the self-identified SZ group (target group) and 257,471 in the CT group. Each text was annotated with post id, redditor, posting data, subreddit, and schizophrenic flair status, being 1 for schizophrenic flair and 0 otherwise.

The target group posting dates range from the first of January 2017 to the end of February 2021, being more concentrated in 2020-2021: Figure 4.15

CT group range from the first of January 2019 to the end of February 2021, being fairly distributed throughout the period: Figure 4.16

Due to data collection strategy and possibly Redditors subject preference, most of the target group data were from the *Schizophrenia* subreddit (87.7%), with only 12,3% from other subreddits 4.17 .

The most representative subreddits of the SZ group after the *Schizophrenia* subreddit were *Ask Reddit*, followed by *No Stupid Questions, Mental Health, Suicide Watch, Neoliberal, Drugs, Am I the Asshole, Unpopular Opinion, 2007 Scape, Relationships, Shower Thoughts, Pokemon Sword and Shield, Depression, Psychosis and Lgbt*, as we can see in figure 4.18.

The subreddits of the CT group included all the subreddits aforedescribed besides Mental Health and Suicidal Watch. The distribution of subreddits in the CT group is

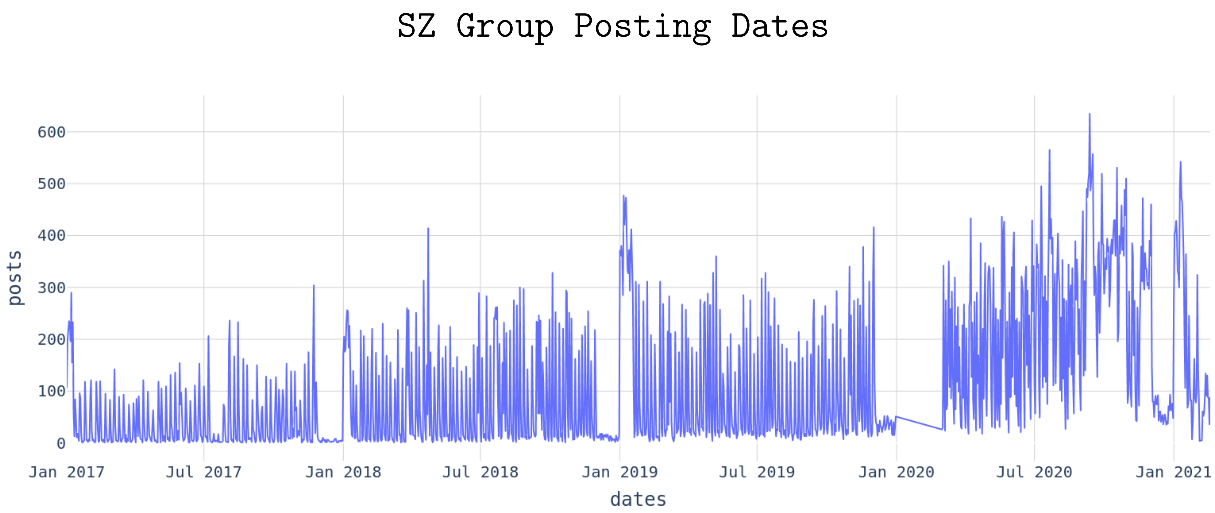


Figure 4.15: Timeline for captured posts for the SZ group

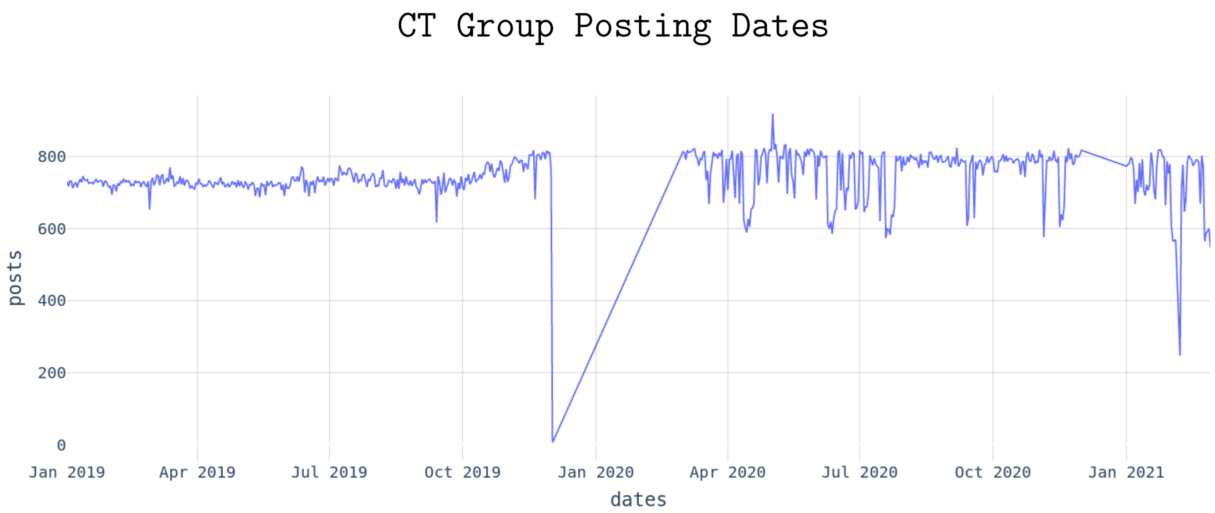


Figure 4.16: Timeline for captured posts for the CT group

illustrated in figure 4.19

SZ group's posts came from 14,667 redditors, each with 10 posts on average. The number of posts per redittors ranged from 1 to 1859. In this group 80% of redditors have 10 or less posts 4.20. On the other hand, the posts from the CT group came from 329,055 redditors, with the average redditor having 1.5 post and only 25% of redditors having more than 1 post 4.21.

Analyzing the word count per text the CT group mean post size was 189 words 4.23 while the target group mean post size was 57 4.22. In the SZ group, 24.5% of the posts were longer than 100 words, while in the CT group a total of 49% of the posts were longer than 10 words.

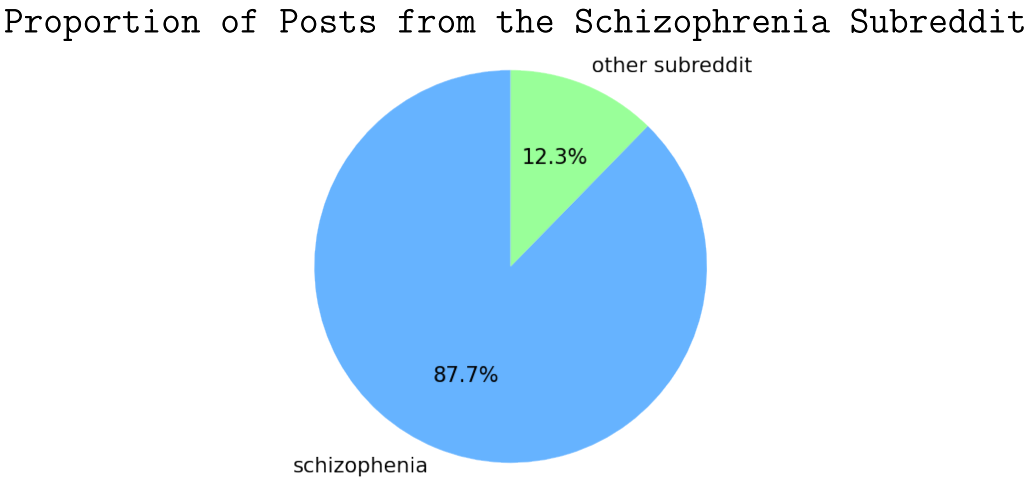


Figure 4.17: Posts from the schizophrenia subreddit percentage

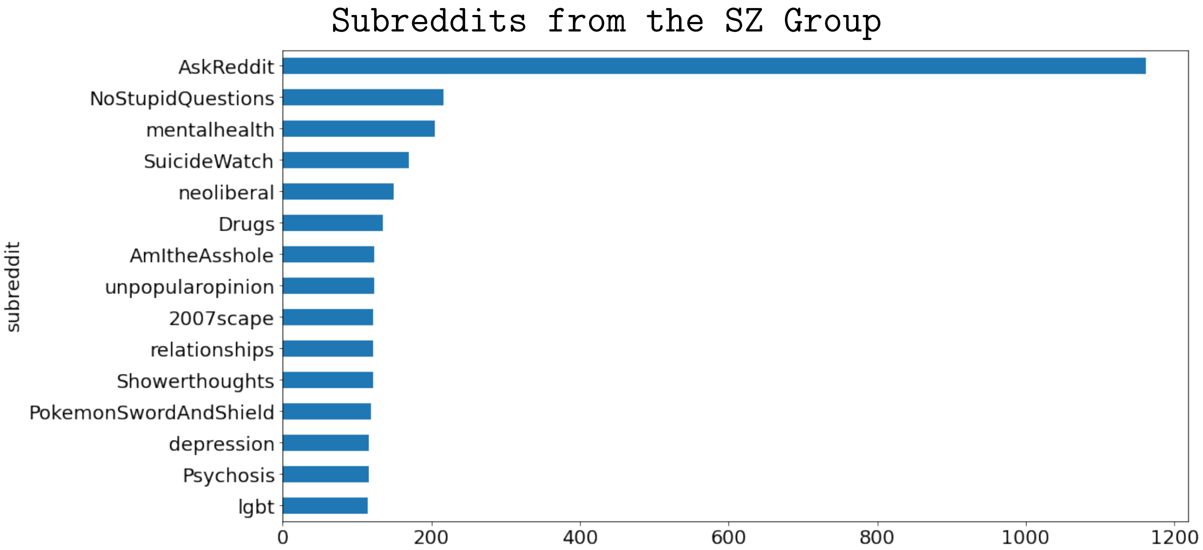


Figure 4.18: Subreddits from the SZ group except the Schizophrenia subreddit

4.2.6 Data Split

In NLP date is a common criterion for dataset split. As language is time sensitive, being that it is constantly changing, training the model on past data to test on posterior data helps build robust models R. One critical attribute of our corpus is the time period, which included the Covid-19 pandemic period, as Covid-19 was declared to be a pandemic by the World Health Organization on 11 March 2020, followed by a quarantine and social distancing period which lasted at least until the end of 2021. Such events impacted people's habits, focus, and mental health R. In this context, since we were present with the opportunity, we decided to split our data on the 11 March 2020 mark, and add another dimension to our analyses. As a result of this data split, SZ group proportion and total

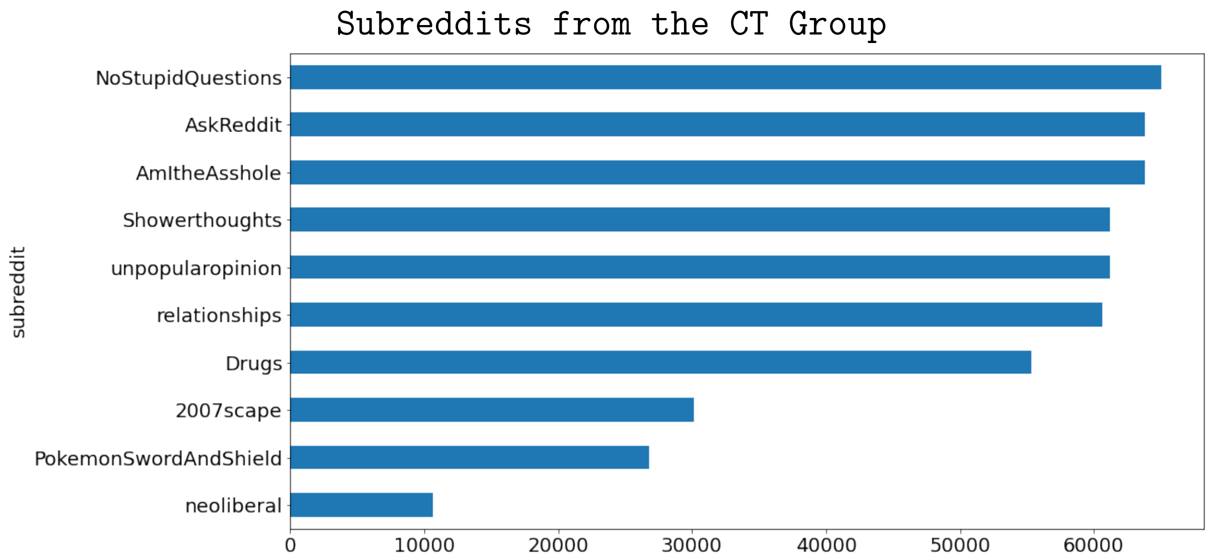


Figure 4.19: All the subreddits from the posts of the CT Group

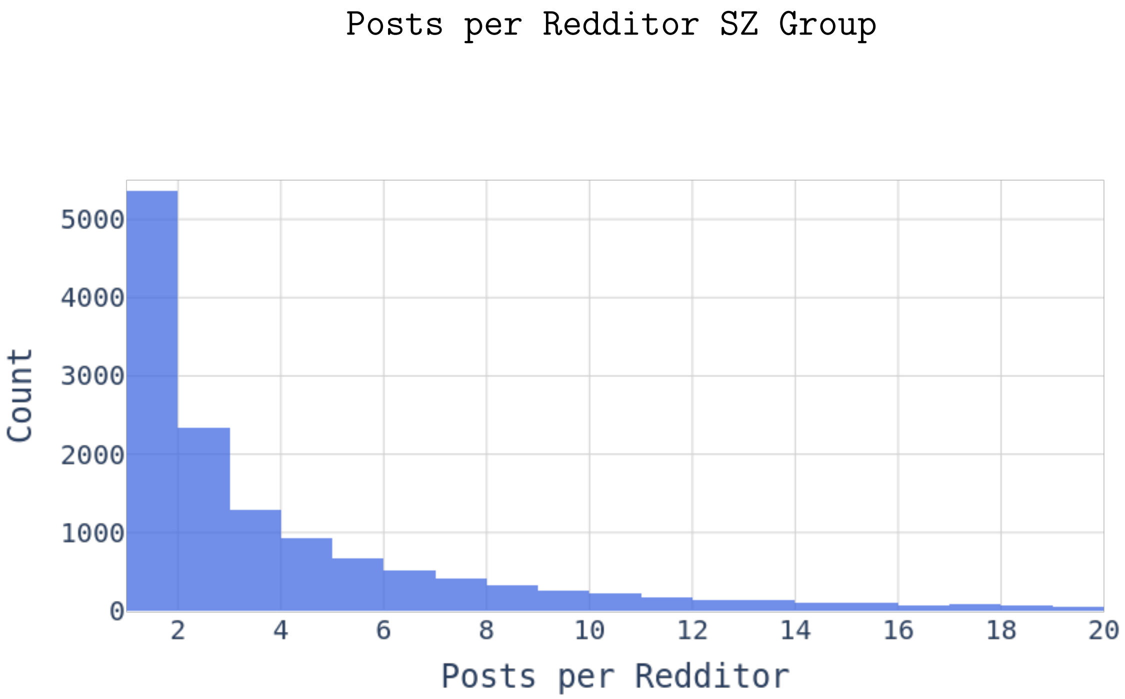


Figure 4.20: Amount of posts for each Redditor in the SZ group

of posts on training and testing datasets are 35.53% of 212,142 and 38.8% of 196,649 respectively.

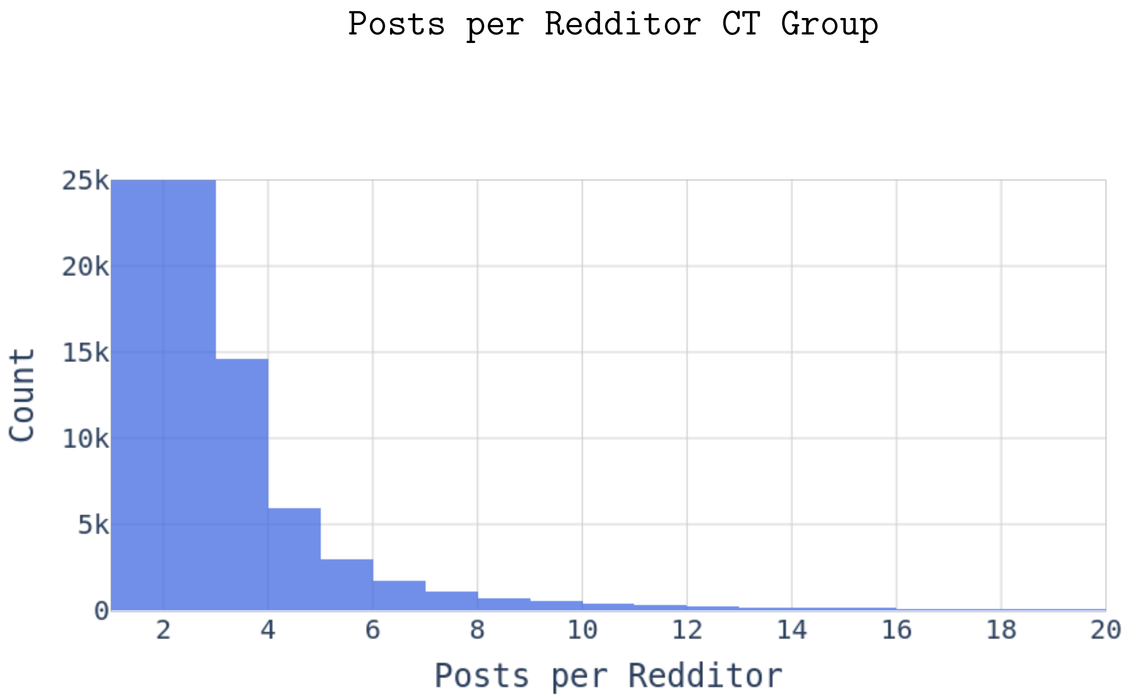


Figure 4.21: Amount of posts for each Redditor in the CT group

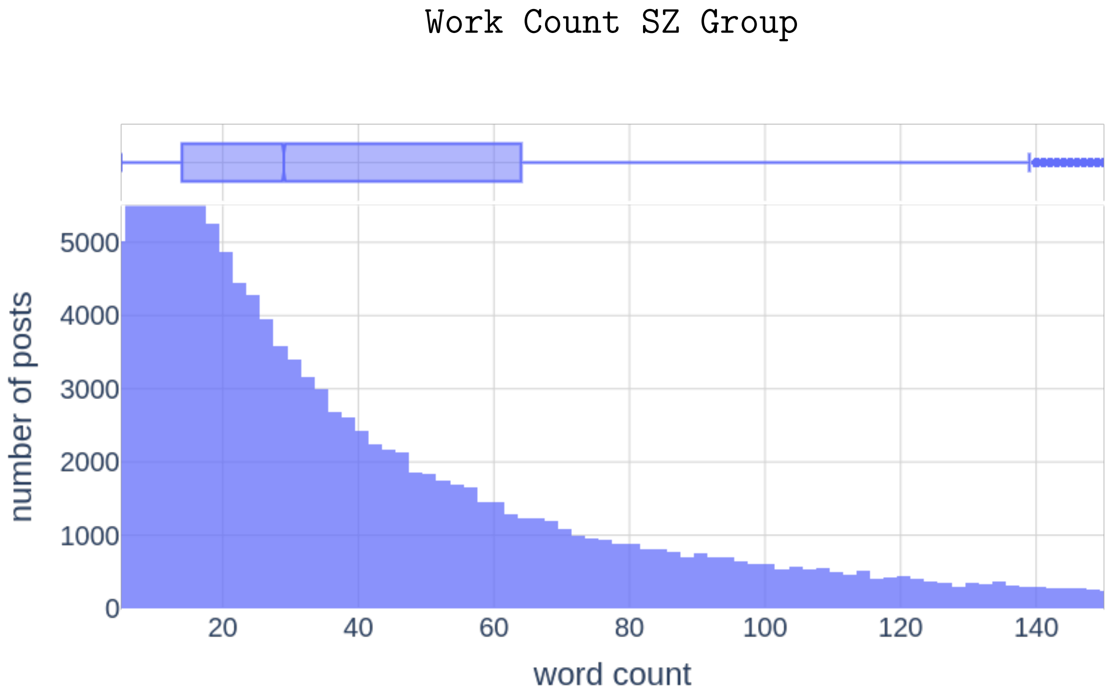


Figure 4.22: Amount of words per post for the SZ group

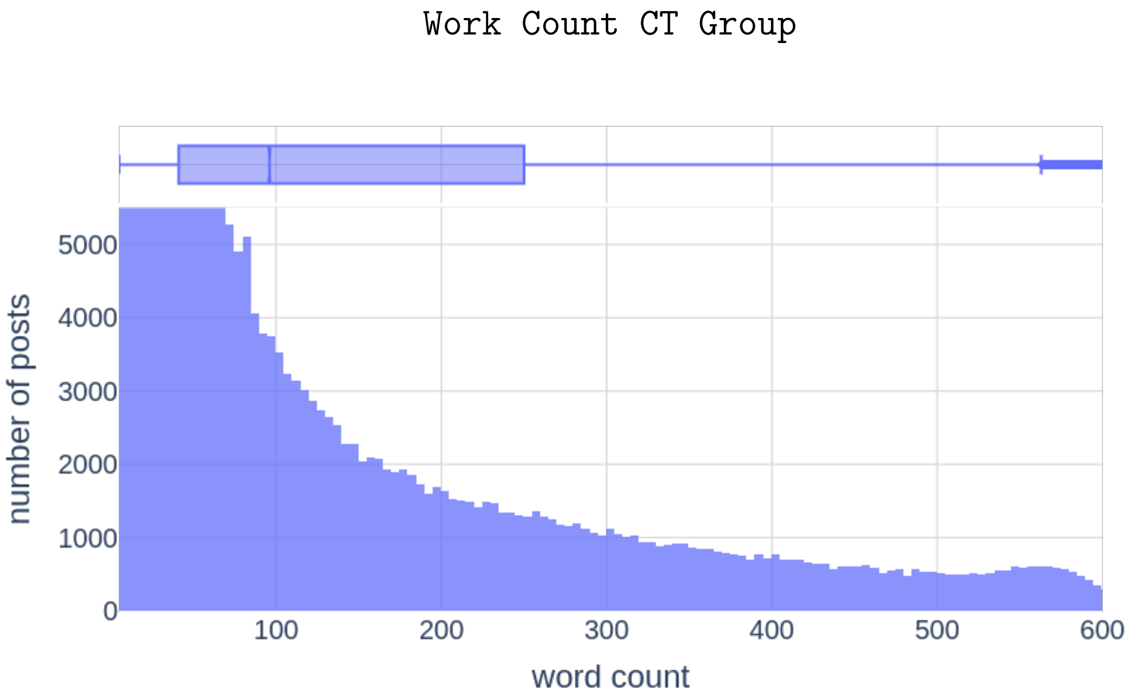


Figure 4.23: Amount of words per post for the CT group

4.3 Metrics

In this section, we present the fundamentals of classification metrics, the most common metrics, and their strengths and limitations. It will be shown how the attributes of the data and the classification proposition were considered in the definition of the chosen metrics of our experiment.

The result of data fed to a binary classifier is one of two classes, usually represented as positive or negative with respect to the target group. That prediction can be either false or true based on the actual label of the input. Hence there are four possible predictions outcomes:

1. FP: false positive (negative wrongly classified as positive)

2. TP: true positive (positive correctly predicted as positive)

3. TN: true negative (negative correctly classified as negative)

4. FN: false negative (positive wrongly classified as negative)

When evaluating a model's prediction of new data inputs, a helpful way to examine the output is through a confusion matrix. It consists of a specific layout with two rows representing positive and negative ground truth labels, and two columns representing positive and negative predictions. From the base values expressed in the confusion matrix, other important metrics can be derived, named:

- Sensitivity: positive class proportion of correctly classified instances
- False Negative Rate (FNR): positive class proportion of incorrectly classified instances
- Specificity: negative class proportion of correctly classified instances
- False Positive Rate (FPR): negative class proportion of incorrectly classified instances
- Accuracy: data proportion of correctly classified instances

However, rather than directly predicting the positive or negative classes, a machine learning model computes the probabilities that an input belongs to one of the given classes (positive, negative). This probability forecast can be interpreted using different thresholds allowing investigation of the trade-off between the TPR and FPR and providing a better understanding of the classifier quality. To that end, a valuable method is to visualize the Receiver Operating Characteristic (ROC) curve, which is the probability curve plotted with sensitivity on the y-axis against the false positive rate on the x-axis. The area under the ROC curve (AUC) represents both classes' degrees of separability. The AUC value is both threshold-invariant and scale-invariant since it measures how correctly the predictions are ranked instead of their absolute values. When AUC is approximately 0.5, it indicates that the model cannot distinguish between positive class and negative class; the higher the AUC, the better the model's discrimination, and an AUC near 0 means the model is reciprocating the result by predicting a class as the other.

An important consideration when the data is imbalanced between classes, especially in the case of this study where the minority class was defined as the positive class, is that the AUC value can be misleading, accentuating the results of the positive class and not allowing a complete estimation of model performance. Nevertheless, the focus on the positive class is often aligned with research objectives, and this effect can be alleviated by carefully paring AUC with other metrics in the analysis to assess the classifier's robustness. This was the case in this study for the feature selection process since it made sense to prioritize precision to maximize the discovery of features that set apart the target group. Later, we considered the F1 and Micro F1-score in addition to AUC in the training process.

The F-score, and other related scores, are metrics calculated from the test's precision and recall. The F1 score is the harmonic mean of precision and recall with equal weights, thus, it assumes equal importance among precision and recall and discourages hugely unequal values and extremely low values. Since we mean to maximize both precision and recall, F1 is a great measure for model evaluation and comparison.

4.4 POS-based features

4.4.1 Part Of Speech Tagging

Part of Speech Tagging (POS Tagging) is a significant task in NLP for syntactic components. This procedure takes into consideration the structural context in which a word occurs in order to assign it a label, a category of the grammatical term. Therefore, a POS tag sequence conveys sentential syntagmatic and paradigmatic relations [57].

In this work, we used SpaCy [26], an open-source Python library that is recognized for achieving, with the employment of deep learning models, state-of-the-art accuracy in tasks such as POS tagging, named entity recognition and dependency parsing. It provides trained components produced with a large set of examples that generalize across a language. Once the component is loaded, it is added to a trained pipeline with the models that make POS tag predictions.

For the POS tagging task, the text is represented in the form of unique numerical values for every input. In this stage, information such as prefix, suffix, and others, are considered in the extraction of values indicating word similarities. The values are fed into a Convolutional Neural Network (CNN) and merged with their context. The result of the encoding process is a vector matrix that reflects the input information. Then, the matrix is passed through an Attention Layer, using a query vector that summarizes the input, and a softmax function predicts the input's part of speech and morphology information. In this process, each word should be assigned with tags of linguistic sets of features.

We resort to Spacy's models for two different types of linguistic annotations: first wide-ranging POS tag such as NOUN (noun), PUNCT (punctuation), ADJ (adjective), ADV (adverb), and others; and second, a fine-grained tag that categorizes a token in more specific categorys. For example, an adjective can be categorized as JJR (comparative adjective), JJS (superlative adjective), or AFX (affix adjective). Every text in the corpus was annotated with tags from the two sets of POS tags, one of each set for every word. It is important to point out that the two types of tag are set independently. This process produced two sequences of tags. We then calculated the occurrence frequency of each tag and normalized the values by dividing them by the total number of words in the text.

4.4.2 Handcrafted Feature Engineering

Analyzing the two sets of tags and taking into consideration recent developments in the understanding of characteristics of language production in schizophrenia and formal thought disorder we designed three new sets of tags. The purpose was to induce automatic recognition and measurement of known grammatical aspects of schizophrenia while creating opportunities for new patterns to emerge.

In designing our experiment, we considered that in order to postulate hypothesis about grammar, it was not essential to follow grammatical taxonomy. However, we did take into account grammar, but it was not considered as a main guideline. Moreover, we were interested in observing the influence of different increasing levels of feature specificity in the results, and for this reason, new tags were devised applying a hierarchical principle. The tag's tree-level structure can be visualized as trees, being the first level tag the root, the second level tags the branches, and the third-level tags the leaves.

We conceived a system of rules that took in the word and the two different spacy-given tags and returned three tags. Even though the system did not receive direct input of the word's sentence, information on word context is present in the spacy-given tags, and therefore, present in the resulting tags. Different combinations and associations of the input tags were used the according to the available tags and motivating hypotheses, and some rules were also based on vocabulary. The tagging system followed the following requirements:

- every word was represented by only one tag in each level
- the word's tags were defined following the order from the first to the third level
- the word's second-level tag only takes into account the choices of tags from the defined first-level tag tree, as the third-level tag only takes into account the choices of tags from the defined second-level tag branch.
- the number of branches and leaves are different in each tagging tree
- if the words second level tag represents a branch that cannot be significatively divided any further the second level tag is duplicated and repeated in the third level.
- If, after the designed rules for filtering the words spacy-given tags were applied, the word did not fall in any pre-established groups for the third level, the words spacy fine-grained tag is added to the branch as a possible third-level feature and designated to the word.

The first-level features represent the main lexical features, namely nominal expressions, adjectives, determiners, verbs, connectors, interjections, pronouns, adjuncts, and particles. For the definition of a word's tag at this level only the spacy's wide-ranging POS tags needed to be considered. This first level's tag established the tree of a word in our hierarchical tagging proposition.

The criteria for the definition of the second-level and third-level features for each tree were different in each case. For example, nominal expressions were branched into proper nouns and generic nouns, and then divided into singular and plural, while pronouns were branched into first, second, and third person, also had branches for singular it, there, interrogative and indicative pronouns and the leaves followed different criteria in each branch. The pronouns tree rules for tagging were heavily based on vocabulary, while the nominal expression tree used only the input tags to select the output tags. Tables 4.1 and 4.2 shows examples of the tagging process results. A complete list and visual representation of each tree's hierarchy can be consulted in appendix A.

Tokens	**1st Level**	**2nd Level**	**3rd Level**	**SpaCy 1st**	**SpaCy 2nd**
Could	1:VERB	2:V_AUX	3:VA_MD	AUX	MD
I	1:PRON	2:P_1st	3:P1_SINGULAR	PRON	PRP
sustain	1:VERB	2:V_VERB	3:VA_VB	VERB	VB
my	1:PRON	2:P_2nd	3:P_2nd(r)	PRON	PRP$
water	1:NOUN	2:N_NOUN	3:NN_SINGULAR	NOUN	NN
needs	1:NOUN	2:N_NOUN	3:NN_PLURAL	NOUN	NNS
based	1:VERB	2:V_VERB	3:VA_VBN	VERB	VBN
solely	1:ADJ	2:A_ADV	3:AAD_DEG	ADV	RB
off	1:ADJ	2:A_ADP	3:AA_IN	ADP	IN
of	1:ADJ	2:A_ADP	3:AA_IN	ADP	IN
salad	1:NOUN	2:N_NOUN	3:NN_SINGULAR	NOUN	NN

Table 4.1: Tagging results first example including all levels of handcrafted tags and the two SpaCy sets of tags that wee used in the created tagging function

Tokens	1st Level	2nd Level	3rd Level	SpaCy 1st	SpaCy 2nd
Hey	1:INTJ	2:I_LEX	3:IL_SOC	INTJ	UH
I	1:PRON	2:P_1st	3:P1_SINGULAR	PRON	PRP
need	1:VERB	2:V_VERB	3:VA_VBP	VERB	VBP
some	1:DET	2:D_DET	3:DD_QNT	DET	DT
advice	1:NOUN	2:N_NOUN	3:NN_SINGULAR	NOUN	NN
on	1:ADJ	2:A_ADP	3:AA_IN	ADP	IN
stomach	1:NOUN	2:N_NOUN	3:NN_SINGULAR	NOUN	NN
tattoo	1:NOUN	2:N_NOUN	3:NN_SINGULAR	NOUN	NN
pain	1:NOUN	2:N_NOUN	3:NN_SINGULAR	NOUN	NN
.	1:INTJ	2:I_NLEX	3:INL_COG	PUNCT	.

Table 4.2: Tagging results second example including all levels of handcrafted tags and the two SpaCy sets of tags that wee used in the created tagging function

4.5 Models Algorithms

4.5.1 Ensemble Learning

A Machine Learning algorithm that takes an ensemble approach relies on the construction of multiple sub-models that together outperform any single one of them. Two popular ensemble methods are bagging and boosting. Bagging uses bootstrap samples from the training set to train several sub-models simultaneously, and each makes a prediction and votes on the prediction to be outputted. Boosting trains sub-models sequentially and each one focuses on improving on the flaws of the last one.

For this experiment, we selected three widely used classification algorithms [20], each from a distinct machine learning method, to assess their classification potential properly and in-depth research contribution, separately and in contrast with each other. To classify hand-crafted features based on part of speech tagging, we opted for the Random Forest and LightGBM algorithms, while for a deep learning approach we employed a transformer model.

Random Forest is an ensemble method that fits several decision tree classifiers, called learners, using a bagging technique and then applying averaging to refine the prediction performance [9]. Each learner is grown from a bootstrap sample of the data using a mechanism in which binary splits recursively partition the inputs. This bootstrap aggregating process, or bagging, decreases the variance of the model without increasing the bias, because although a single learner is sensitive to noise, each one is provided a distinct training set, resulting in a group of learners with no correlation that together produces a better classifier. Besides applying bagging to the data sample each learner will see, the features each learner will consider are also selected by a bagging strategy. This contributes to lower variance, avoiding the same features being chosen by several learners and the subsequent learners' correlation. Consequently, the most relevant random forest hyperparameters are the number of learners the algorithm builds, the maximum number of features a learner considers, the minimum size of sample required for a split, and the minimum sample required for each partition.

We should note that when the data is very sparse, it's possible that for some splits, the bootstrapped sample and the random subset of features will collaborate to produce an invariant feature space creating unhelpful learners. This was a risk with our data and the feature engineering approach we have taken and required attention that included feature selection and data normalization.

LightGBM is a gradient boosting framework that also uses tree-based learning algorithms. Differently from Random Forests, each learner in LightGBM minimizes a loss function by iteratively choosing a function that means to improve the negative gradient from the previous learners [27]. Once the learners are trained to correct each other's errors, they are expected to be more capable of capturing complex patterns in the data. The LightGBM algorithm utilizes the Gradient-Based One-Side Sampling (GOSS) method along with the Exclusive Feature Bundling (EFB) technique. GOSS focuses on retaining the data with a large gradient, which needs to be improved to enhance model performance, and does random one-side sampling on data with a small gradient, resulting in a reduced search space and faster training. EFB is a method that reduces the number of effective features by bundling nearly exclusive features into a single feature, reducing dimensionality and better handling sparse feature space.

Even though Gradient boosting trees have shown to be more accurate than random forests, if there is much noise in the data, the boosted trees may overfit and start modeling the noise. To suppress this effect, feature selection can help reduce noise by excluding irrelevant or redundant features, additionally, controlling the minimum data size for each leaf can prevent the model from becoming too specific. Important hyperparameters for a LightGBM model are the maximum number of bins that feature values will be bundled in, the tree learners' depth limit, and minimal data size in one leaf.

4.5.2 Transformer

To establish a baseline for the classification task, we implemented a transformer model using the transformers package from Huggingface [60], an open library and centralized platform that provides a common API for easily loading the weights of transformer-based models with different frameworks and original code bases.

While we were carrying out this experiment, the transformer was the state of art architecture for natural language processing due to its architecture scalability with training data that captures long-range sequence features with facilitated parallel training. The huggingface library provided the implementation of both the tokenizer and language model. We chose to employ the DistilBertTokenizer, a tokenizer with a 30,522 tokens vocabulary, and with pre-trained weights of the distilBERT[43] language model that were retrieved from the checkpoint 'distilbert-baseuncased'. Tokenizer classes store the vocabulary token-to-index map for their corresponding model and handle the encoding and decoding of input sequences according to a model's specific tokenization process. Tokenizers can also implement additional useful features as token type indices and maximum length sequences.

We trained the transformer model using the *DistilBertForSequenceClassification* model architecture, an effective less computationally intensive alternative to BERT. The backend used for training was PyTorch, and we also took advantage of resources like the Scikit-learn library and Google Colab. The transformer model had six hidden layers and twelve attention heads. The maximum input size was set at 512. The initial learning rate was set to 0.001 and is reduced by a factor of 0.95 if the validation loss did not decrease for three epochs. The model was fitted with a batch size of 32 and trained for four epochs. The model configuration is described in Table 4.3.

4.6 Feature Selection

Feature selection is the process of reducing the number of input variables when developing a model by automatically or manually selecting those features which contribute most to the improvement of model performance [56]. Reducing the number of input variables is desirable to reduce the computational cost of modeling and the training time. It can also be a crucial step in reducing the risk of overfitting since less redundant data means less opportunity to make decisions based on noise.

activation function in the encoder and pooler	gelu
attention dropout ratio	0.1
Dimensionality of the encoder layers and the pooler layer	768
dropout for all fully connected layers	0.1
size of the intermediate enconder layer	3072
initializer range for weight matrices	0.02
maximal position embeddings	512
Number of attention heads for each attention layer	12
Number of encoder hidden layers	6
sequence classification dropout	0.2
Vocabulary size of the DistilBERT model.	30522

Table 4.3: Trasfomer Model Configuration

In this work, feature selection had another important role since we were creating a model to distinguish between posts of people with schizophrenia and those in the control group, not only to automatically identify the correct group but also to understand how they differ and, more specifically, the characteristics of the schizophrenic language that can be used as a biomarker.

Feature selection was an essential part of assessing the impact of the level of specificity of the POS tag used. We were interested in the ranking of the selected features, the feature's contribution, comparing features selected at different levels, and the number of features for best performance, among other analyses.

Considering that we wanted the features to highlight the patterns of the target group as much as possible, we chose a supervised approach, selecting features based on the target class and using AUC as a parameter. The supervised approach that better fitted our objective was a method that searches the space of possible subsets of features, assessing their quality by learning and evaluating a classifier with that feature subset based on the same machine learning algorithm used for training. This greedy search approach is referred to as a wrapper technique.

We selected the features first by calculating the AUCs of single feature classifiers, and then we performed a forward selection by selecting the second variable that best performs in combination with the first one. This process continued until we ranked all the variables and registered the AUC increase of each feature added.

To get an accurate estimate of the classifier's error on new data, we used a cross-validation technique named One-Group-Out. We defined ten groups by sorting groups of Redditors and attributing all their posts to the same group. We used the last letter of the id of the users as a criterion to separate the groups. Therefore to test each feature to be added to the selection, there were ten iterations of classification, each of which treated nine groups as the test set and trained on the other group. Once all ten iterations were completed, the resulting iterations were averaged together, creating the final cross-validation model.

This process was carried out until the last feature was ranked. That is not usually necessary, but as we mentioned before, we were aiming to get further insight into the feature's ranking. We decided to use the same number of features in the second, the third, and the combined set of features. Since the first level was considerably smaller, it was clear that its size could not be set as a parameter. We could have stopped including features when they did not add a significant increase in performance. But this approach might not provide enough features for the comparison analysis we intended.

Instead of defining beforehand the number of features, we decided first to rank all features and then choose by analyzing the curves of the AUC increase of the 4 sets of features, the limit of features that could provide both performance and analytical insight.

Chapter 5

Experiment

5.1 Baseline

In the first attempt at modeling using our transformer architecture setup both training and test metrics were promising. With a test set weighted average F1 score of over 0,96 , the transformer model could be considered successful. Analysis of the layer-integrated gradients explanations of the results indicated that the model was making predictions mostly based on SZ related vocabulary. In table 5.1 we can see a list of the most relevant words in the correct predictions of the SZ group. The high weight of words such as disorder, medicine, psychiatrist, disability, psychotic, diagnosed, and schizophrenia, in this results, suggested that the patterns found by the model were topic related with no indication of syntactic aspects being considerate.

Considering our corpus characteristics we found the explanations showed a strong indication that the model might be recognizing not the texts of people with SZ, but the texts with the topic of SZ, more specifically, the texts from the subreddit *Schizophrenia* subreddit . This is an important distinction that cannot be ignored. Even though we did expect or establish that the baseline model should find syntactic patterns, if it used the data label to train in what seems to be a different task, it could not serve as a baseline.

We then moved to the second attempt, in which we removed from the data all posts from the schizophrenia subreddit and trained the model with the new data. The new model achieved an F1 score of 0.91. In table 5.2 can see a list of the most relevant words in the correct predictions of the target group in this second phase. Since the second trained model demonstrated less bias we will consider it as the baseline in this experiment.

Word	**Contribution**
disorder	0.1691
shower	0.1717
medicine	0.1725
art	0.1734
psychiatrist	0.1750
disability	0.1803
mood	0.1810
humans	0.1864
psychotic	0.1869
!	0.1906
hugs	0.1909
yes	0.1953
psycho	0.2002
illusion	0.2022
therapy	0.2093
symptoms	0.2100
voices	0.2128
oh	0.2191
earth	0.2257
thanks	0.2305
diagnosed	0.2341
med	0.2358
technically	0.2473
schizophrenia	0.2536
hospital	0.2693
thank	0.2877
diagnosis	0.2912

Table 5.1: Most Relevant Words in Correct Predictions of the Target Group – first attempt

Word	**Contribution**
heck	0.2304
description	0.2333
color	0.2349
blood	0.2351
piercing	0.2377
oh	0.2445
pen	0.2475
edition	0.2615
pc	0.2627
cards	0.2637
syndrome	0.2681
audio	0.2703
yes	0.2712
readings	0.27436
cd	0.2778
álbum	0.2848
protein	0.2884
yeah	0.2917
sims	0.3023
settings	0.3038
psychic	0.3138
diagnosis	0.3185
ink	0.3258
lip	0.3260
pup	0.3510
card	0.3694

Table 5.2: Most Relevant Words in Correct Predictions of the Target Group – second attempt

5.2 Selected Features

The feature selection process produced similar curves of AUC increase with each addition of features for the third level set of features and one including all the 3 levels, as we can observe both in the LGBM models (5.3) and Randon Forest models (5.4). The model's performance rank was for both algorithms: first-level model, second-level model, three levels model, and third-level model. There was little increase in performance after the 12th feature of the 2nd level models and the 20th feature of the 3rd and 3 levels model. Taking these results into account, we decided to establish the maximum number of features at 20.

Besides reducing the dimensionality of the models, the feature selection procedure revealed the rank of selected features. The first feature of each LGBM model and its respective AUC was: noun in the 1st level with 0.669 score; first person pronoun in the 2nd level with 0.639 score; first person pronoun singular in the 3rd level with 0.650 score; and noun in the three levels with 0.669 score. It was not surprising that two of the chosen features were in the same branch. Even though it was by design that the first feature of the model with three levels would be the same as the 1st feature from the other models with the higher score, it is interesting that this feature belonged to the level of least specificity. That indicates the heavy influence of noun tree features in distinguishing the texts produced by people with schizophrenia. In the 1st and 3rd Random Forest models, the first feature was the same as the LGBM models at the same level of granularity, and in the 2nd level model, the first feature was lexical interjections while the first person pronoun was the second. The Random Forest model which included all features, started with the first person pronoun singular, which is a feature from the most specific level. Nevertheless, this feature was the 4th from the rank of the LGBM model.

Both models trained with 1st level features ranked the features in the same order. Comparing models that took the same set of features but with different algorithms, even though there might be some distinctions, they are mostly minor differences in rank, and there are at most two exclusive features in each. Features from the noun, pronoun, and interjection trees were highly ranked, as the ones related to the 1st person were more prominent, followed by the 2nd person and then the 3rd person. The features related to singular forms outstood plural forms.

We explored the correlation among the selected features from both algorithms, comparing the schizophrenia and the control group. In the first level, the verb was the feature with more extreme correlations for the target group, while in the control group, it was noun. Connectors showed less correlation in the control group, being a feature that is equally present in both groups. Figure shows the contrast of the percentual of first-level features occurrence between the two groups.

Interjection did not elicit much attention considering the occurrence amount or extreme correlations, nevertheless, it was the second feature to be selected for both algorithms. It was also a top feature for the LGBM 2nd level model, in that case, non-lexical interjections were elected in the 4th place, and lexical interjection was later designated in the 8th place. The Random Forest algorithm ranked non-lexical and lexical interjections in the first five features.

For the 3rd level, the LGBM and Random Forest models appointed first the social lexical interjections, and the non-lexical interjections, with the Random Forest model selecting the emotional interjection before the cognitive, and the LGBM selecting these features in reverse order. When considering features from all levels Interjection was one of the only two features from the first level to be selected, indicating that the entire set of linguistic elements represented in the root of the interjection tree was more relevant together than it was separated. Also, for the selection with the 3 sets of features, the non-lexical cognitive feature was selected, meaning that it still added value even after the first level feature was included. The Random Forest algorithm, on the other hand, did not select root features when provided with features from all the levels. In fact, this is an important difference between selected features when we compare the feature selection results of both algorithms, and we will discuss this further in our analysis. Nevertheless, the features from the interjection tree of the second and third levels were selected.

The fact that interjection had such role in classification, even with not much prominence when constrasting groups occurrences, highlights how this approach can help bring forward some relevant linguistic characteristics that are not so distinguishable in frequency analysis or correlation analysis. On that note, it is paramount to keep in mind that features were selected by the increase in AUC when they were added. This could point to a series of causes, such as they helped better classify inputs that were already with a score close to the threshold of the correct class, or they brought out a niche of inputs with a distinct group of characteristics, to list a few. Some features with a high correlation with previously selected features, might also represent an important characteristic in discerning among classes but weren't selected because they did not add much to what was already being correctly classified. Thus frequency and correlation analysis can be complementary to this work's machine learning approach and can play a role in formulating hypotheses and deriving highlights from the results.

When considering the order of the features, we can see in figure that the leading 8 to 9 features bring the most prominent rise to the cumulative AUC curve. Even though the feature selection order can be thought-provoking and provide insight into the feature's discriminative effect, the conclusions we can derive from them are limited. The model takes no consideration of the order of the features, and their association might take a different basis. On account of this, feature final contribution might be better accessed by adding explanation techniques to the analysis. However, prior to explainability, we will contemplate test results on data not used for training and evaluate the model's generalization.

rank	LGBM 1st Level	LGBM 2nd Level	LGBM 3rd Level	LGBM All Levels
1	1:NOUN: 0.669	2:P_1st: 0.639	3:P1_SINGULAR: 0.650	1:NOUN: 0.669
2	1:INTJ: 0.692	2:P_2nd: 0.714	3:P_2nd(r): 0.724	2:P_INT: 0.733
3	1:CONN: 0.706	2:N_PROPN: 0.750	3:NP_SINGULAR: 0.759	3:AAD_WH: 0.757
4	1:VERB: 0.714	2:I_NLEX: 0.773	3:P_INT(r): 0.780	3:P1_SINGULAR: 0.778
5	1:PRON: 0.722	2:P_INT: 0.791	3:AAD_WH: 0.793	3:P3_SINGULAR: 0.791
6	1:DET: 0.728	2:N_NOUN: 0.802	3:P3_SINGULAR: 0.806	1:INTJ: 0.802
7	1:PART: 0.733	2:P_3rd-it: 0.810	3:NN_SINGULAR: 0.817	2:P_2nd: 0.815
8	1:ADJ: 0.733	2:I_LEX: 0.818	3:IL_SOC: 0.825	2:N_PROPN: 0.827
9		2:D_NUM(r): 0.824	2:DN_NUM(r): 0.831	2:D_NUM(r): 0.832
10		2:P_it: 0.827	3:INL_COG: 0.837	3:CS_COND: 0.836
11		2:V_VERB: 0.829	3:INL_EMOT: 0.843	3:INL_COG: 0.840
12		2:V_AUX: 0.831	3:CS_COND: 0.847	3:VA_VB: 0.843
13		2:D_DET: 0.833	3:DD_DEM: 0.849	3:DD_DEM: 0.846
14		2:P_IND: 0.833	3:P_it(r): 0.852	3:DD_POSS: 0.848
15		2:C_CCONJ: 0.834	3:DD_POSS: 0.855	2:P_it: 0.850
16		2:A_ADJ: 0.835	3:VA_VB: 0.857	3:NN_PLURAL: 0.853
17		2:A_ADP: 0.836	3:P1_PLURAL: 0.859	3:VA_VBG: 0.855
18		2:C_SCONJ: 0.836	3:VA_VBG: 0.860	2:V_AUX: 0.856
19		2:A_ADV: 0.837	3:VA_VBP: 0.861	3:P1_PLURAL: 0.858
20		2:PT_PART(r): 0.837	3:NN_PLURAL: 0.862	3:VA_VBP: 0.859

Table 5.3: Rank of the top 20 features selected for each of the LGBM models, with the accumulated AUC value

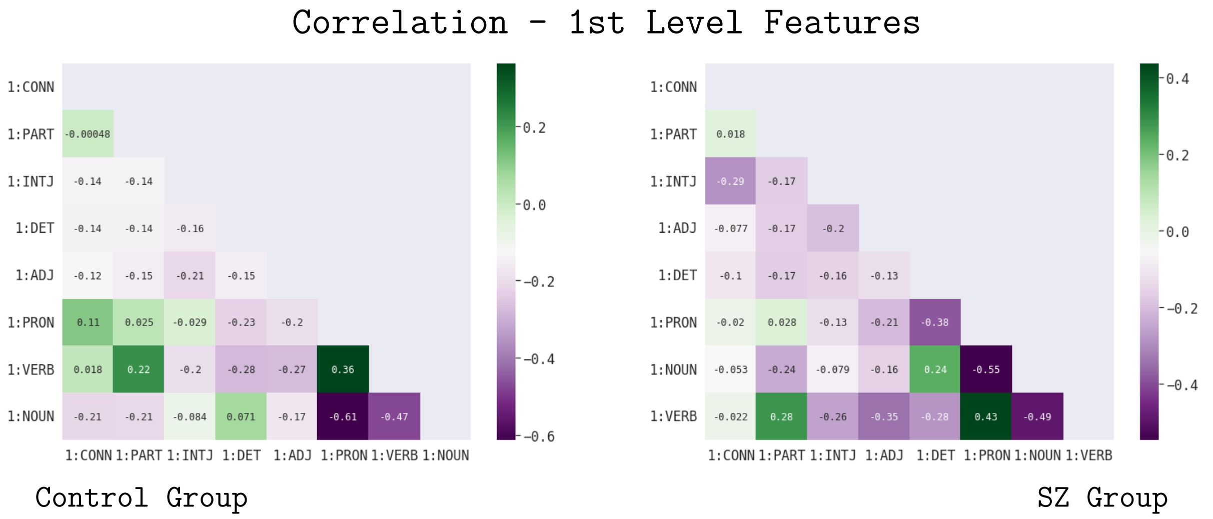


Figure 5.1: Pearson's Correlation Coefficient Matrices of 1st Level Features for the SZ e Control Group

rank	RForest 1st L	RForest 2nd L	RForest 3rd L	RForest All L
1	1:NOUN: 0.554	2:I_LEX: 0.543	3:P1_SINGULAR: 0.555	3:P1_SINGULAR: 0.555
2	1:INTJ: 0.569	2:P_1st: 0.573	3:NP_SINGULAR: 0.598	2:N_PROPN: 0.601
3	1:CONN: 0.575	2:P_2nd: 0.610	2:DN_NUM(r): 0.621	2:DN_NUM(r): 0.627
4	1:VERB: 0.580	2:N_PROPN: 0.625	3:IL_SOC: 0.641	3:IL_SOC: 0.642
5	1:PRON: 0.584	2:I_NLEX: 0.642	3:AAD_WH: 0.657	3:P3_SINGULAR: 0.657
6	1:DET: 0.585	2:P_it: 0.649	3:P_2nd(r): 0.666	3:AAD_WH: 0.667
7	1:PART: 0.587	2:D_NUM(r): 0.655	3:P_INT(r): 0.674	2:P_2nd: 0.674
8	1:ADJ: 0.590	2:P_INT: 0.660	3:P3_SINGULAR: 0.680	2:P_INT: 0.681
9		2:N_NOUN: 0.666	3:CS_COND: 0.684	2:I_NLEX: 0.685
10		2:V_VERB: 0.671	3:INL_EMOT: 0.690	3:NN_SINGULAR: 0.690
11		2:V_AUX: 0.674	3:NN_SINGULAR: 0.692	3:CS_COND: 0.695
12		2:P_3rd-it: 0.676	3:VA_VB: 0.696	3:VA_VB: 0.699
13		2:D_DET: 0.677	3:INL_COG: 0.700	3:DD_DEM: 0.701
14		2:C_CCONJ: 0.677	3:DD_POSS: 0.703	2:P_it: 0.704
15		2:A_ADV: 0.678	3:DD_DEM: 0.703	3:INL_EMOT: 0.705
16		2:C_SCONJ: 0.679	3:AAD_DEG: 0.704	3:AAD_DEG: 0.707
17		2:P_IND: 0.679	3:P_it(r): 0.709	2:V_AUX: 0.708
18		2:PT_PART(r): 0.679	3:VA_VBD: 0.711	3:DD_POSS: 0.711
19		2:A_ADJ: 0.679	3:P1_PLURAL: 0.712	3:VA_VBZ: 0.712
20		2:P_there: 0.678	3:VA_VBG: 0.712	3:VA_VBD: 0.712

Table 5.4: Rank of the top 20 features selected for each of the Random Forest models, with the accumulated AUC value

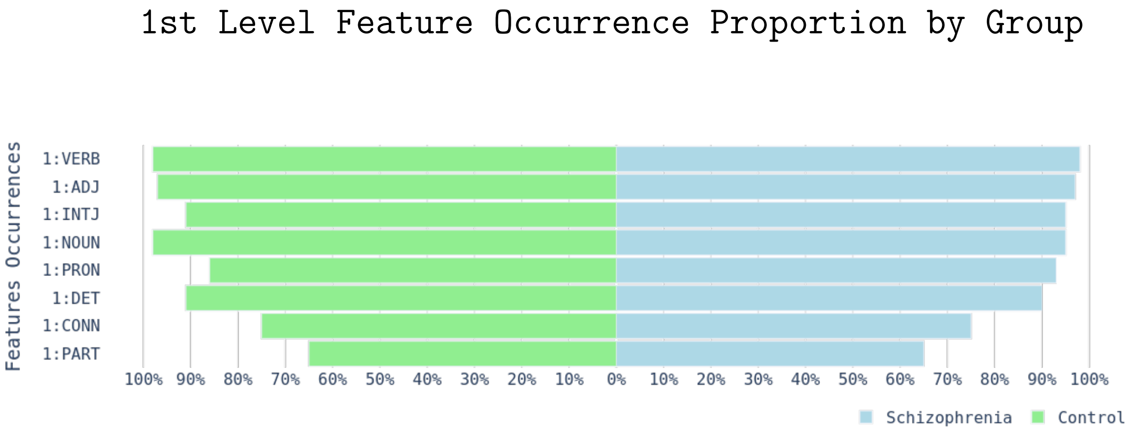


Figure 5.2: Percentage of feature occurrence 1std Level Features for the SZ and Control Group

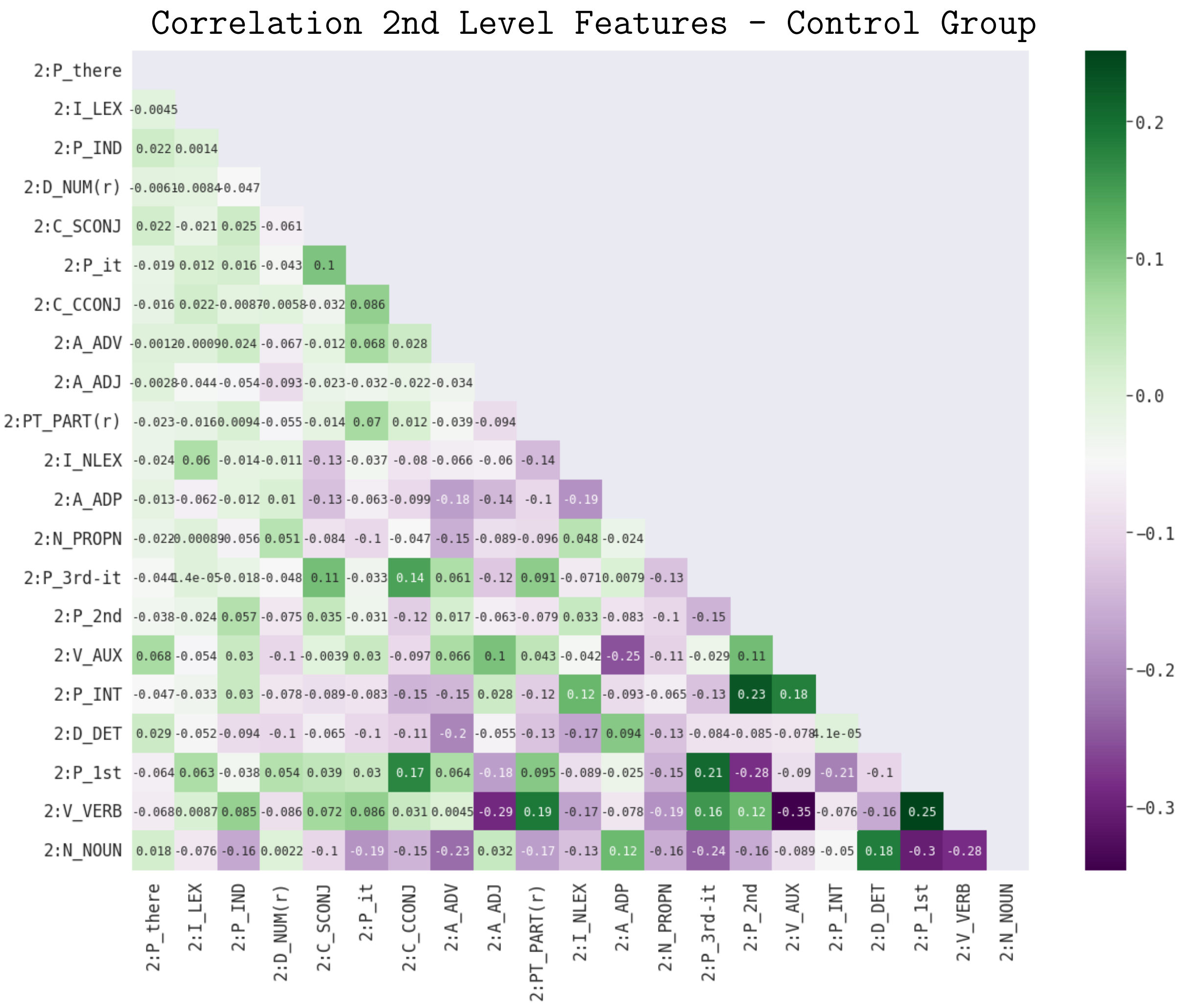


Figure 5.3: Pearson's Correlation Coefficient Matrices of 2nd Level Features for the Control Group

5.3 Test Results

The metrics on the train and test results are displayed in the tables 5.12, 5.13 and 5.14. We used the same hiperparametres for all models with the same algorithm for better comparison. Better results can be achieved by applying optimization methods. Overall, the models trained with the LGBM algorithm achieved the best results with all the feature sets. The LGBM model trained with the 3rd level of features achieved the best results in training, followed closely by the LGBM model trained with features from all the levels. Nevertheless, the F1 results of the Random Forest models were close to those of the LBMG models.

As mentioned before, we used a cross-validation technique to estimate model generalization when training the models in order to assess how the models will perform to an unseen dataset. However, is the distance between the training and test results that

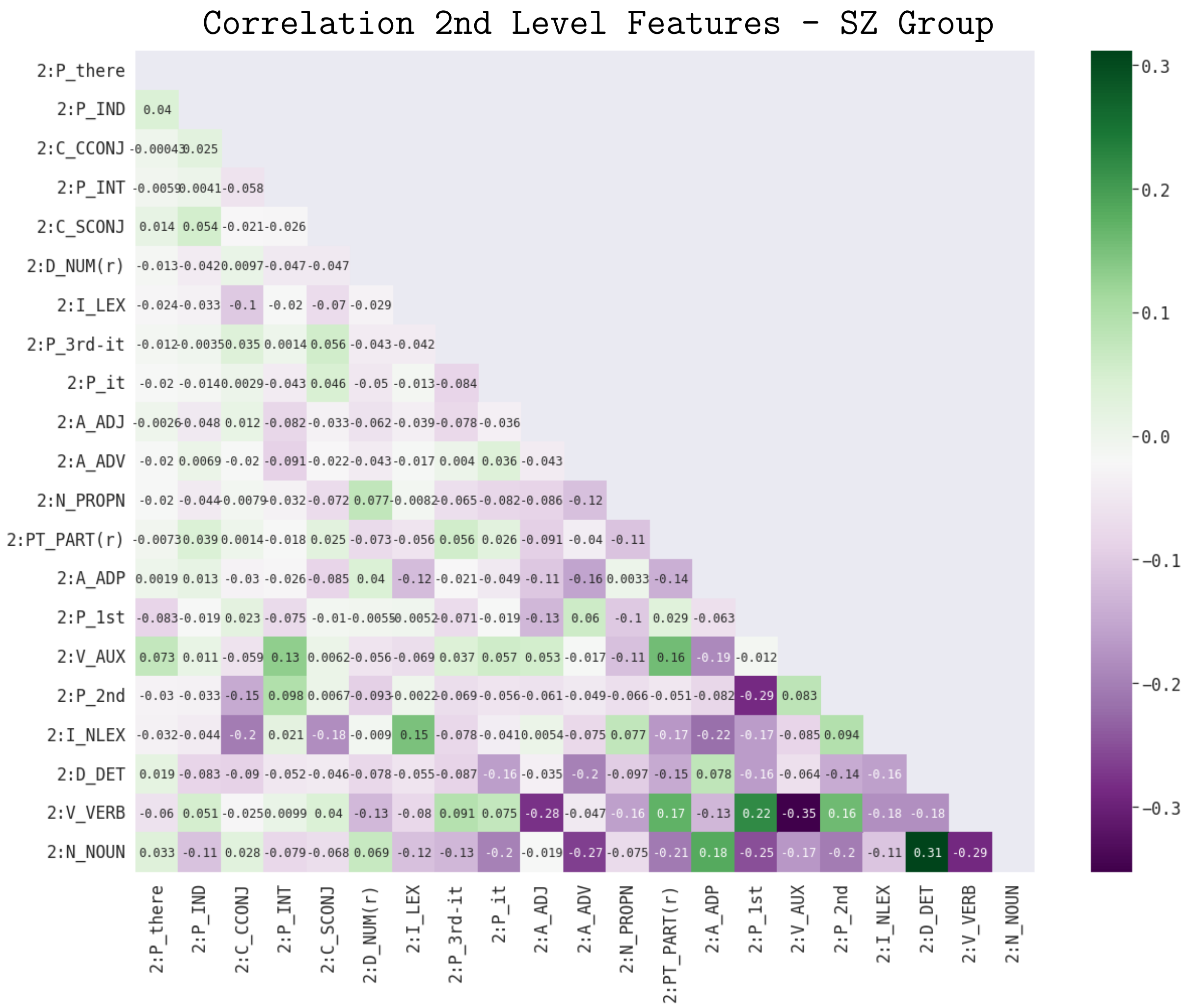


Figure 5.4: Pearson's Correlation Coefficient Matrices of 2nd Level Features for the SZ Group

really provides an insight in whether the model is overfitting the data. In image N we can observe that even though the LGBM 3rd Level model performed better in training, it was the LGBM model with all the levels that achieved the test's best results. More importantly, the difference among the training estimation and the test results was the lowest for the LGBM model with all feature levels, witch indicates it is the most robust model. A similar effect can be noted in the Random Forest models results, with the model with features from all levels having the best approximation generalization balance.

Since the training and the testing datasets were split at the data of the declaration of the covid-19 pandemic, we could address whether a model trained in this classification task could still be applied in such an abrupt change of conditions. The stress from isolation and social disruption during lockdown periods might aggravate SZ symptoms, and an increase in anxiety and depression, could generate noise in the data. The reported minor model estimation error provides insight to this matter, pointing to the model still being effective in this situation. An explanation for this finding can be that the symptoms

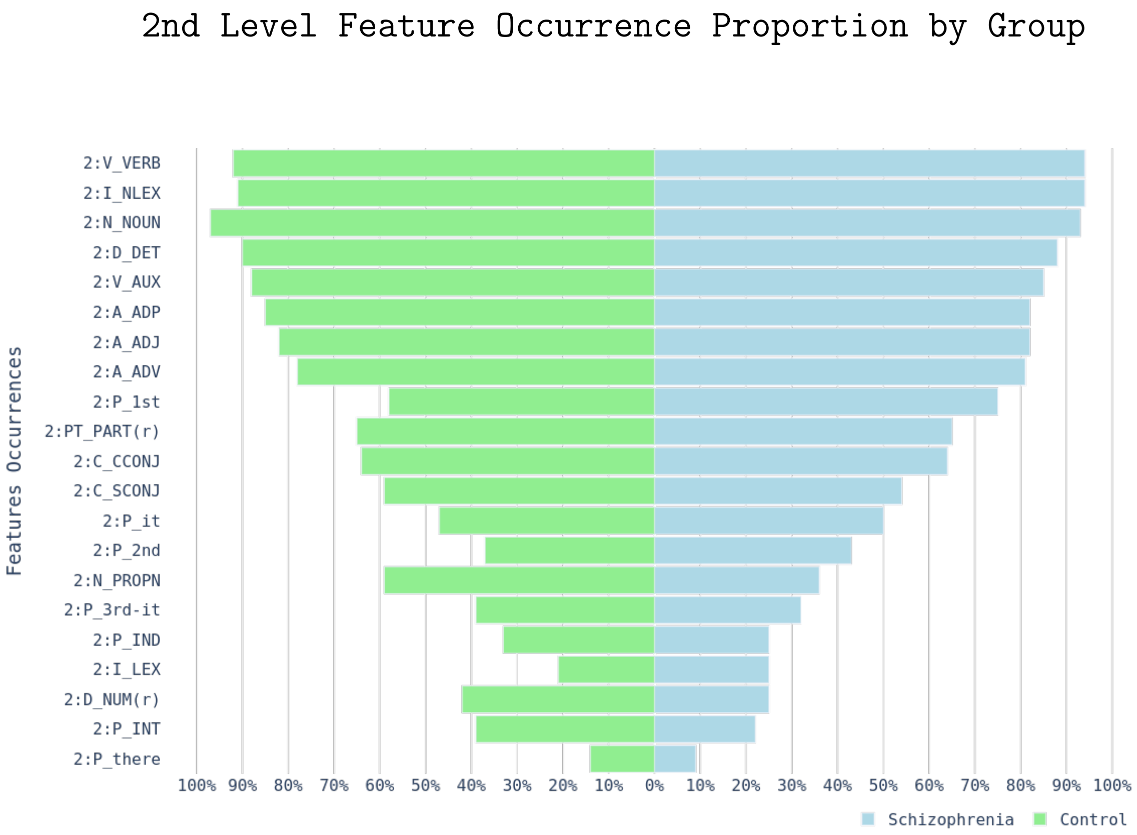


Figure 5.5: Percentage of feature occurrence 2nd Level Features for the SZ and Control Group

might vary in severity and affect the linguistic anomalies' frequency or intensity, but the did not inherently change the linguistic makers' presentation [53].

model	auc	f1 micro	f1 macro	f1_weighted
LGBM 1st Level	0.727	0.776	0.560	0.721
LGBM 2nd Level	0.823	0.815	0.697	0.796
LGBM 3rd Level	0.846	0.830	0.731	0.816
LGBM All Levels	0.846	0.830	0.732	0.816
Random Forest 1st Level	0.577	0.766	0.582	0.727
Random Forest 2nd Level	0.672	0.813	0.696	0.795
Random Forest 3rd Level	0.703	0.828	0.730	0.815
Random Forest All Levels	0.705	0.829	0.730	0.815

Table 5.5: Models Test Results

Analyzing the Kappa Coefficient among models, illustrated in figure 5.15 we can see that the models with the 3rd level features and the model with features from every levels trained with the same algorithm show the highest agreement. The model that showed better agreement with models from the other algorithm was the LGBM 3rd level

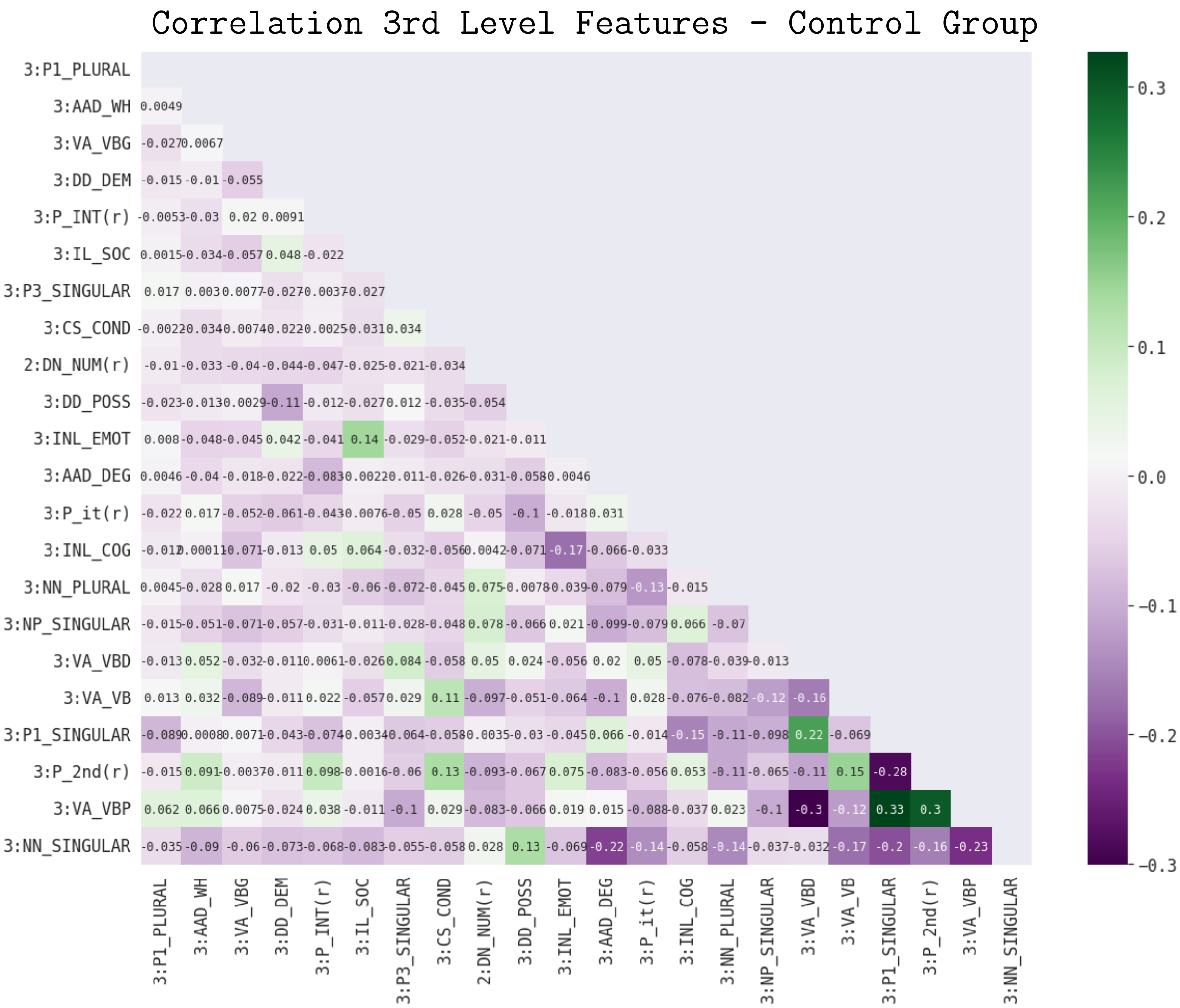


Figure 5.6: Pearson's Correlation Coefficient Matrices of 3rd Level Features for the Control Group

model. The model with the lowest overall agreement with other models was the 1st Level LGBM model. Even thought both 1st level models ranked the features in the same other we can see the two of them had the lowest Kappa coefficient.

5.4 Explanations

The global explanations of the best performing models trained with each dataset are represented in the summary plots in figures 5.16. Each point on a plot reflects the calculated feature shapley value in a respective input, and the color indicates the input value of the feature. By analyzing the distribution and color of the points for a feature, we can understand the impact of feature value on the model prediction and derive insights

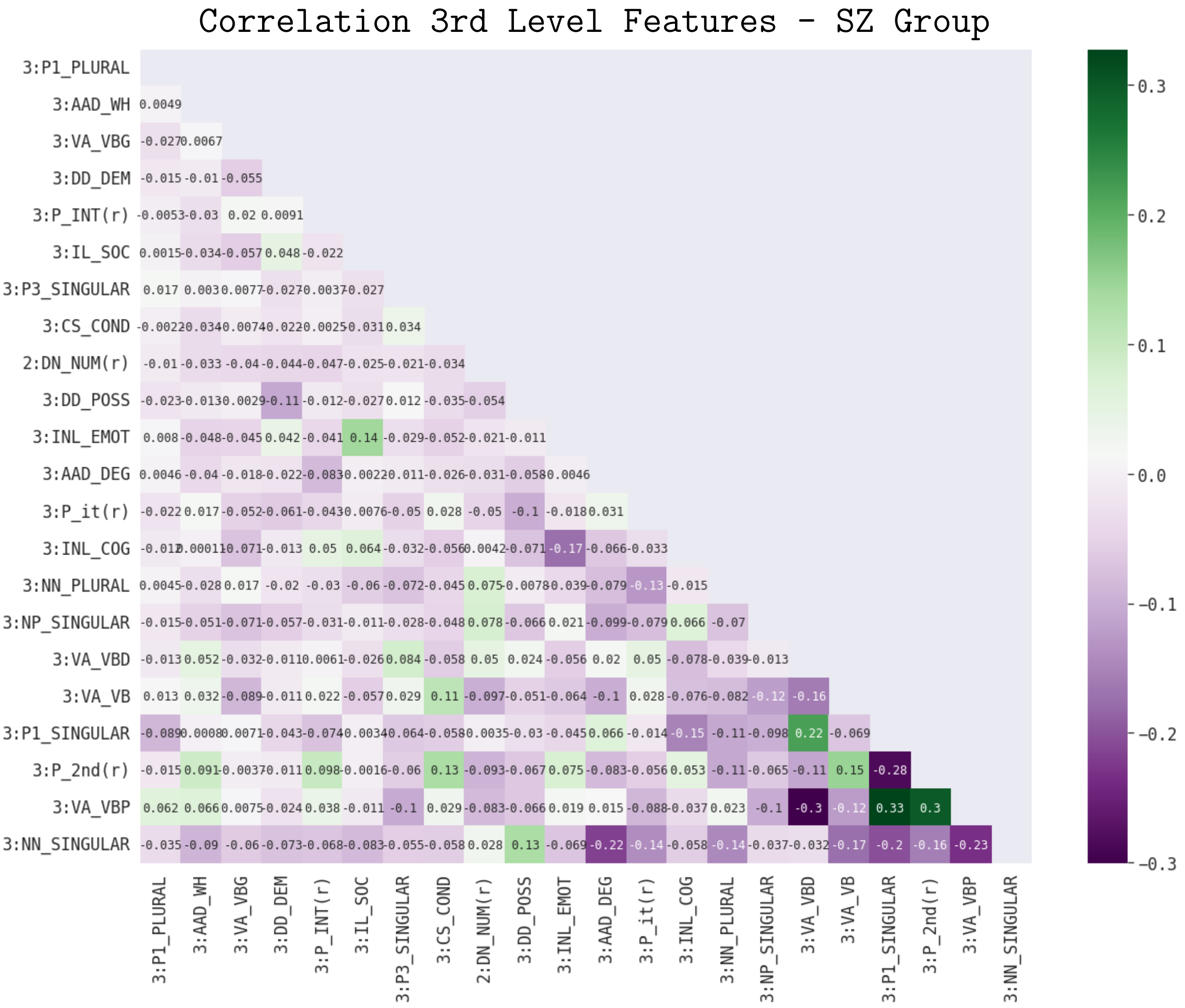


Figure 5.7: Pearson's Correlation Coefficient Matrices of 3rd Level Features for the SZ Group

from the modeled language patterns.

Our findings are consistent with previous studies analyzing the syntactical language disturbance in SZ particular with those working with written texts. It has been demonstrated that pronouns are a prominent marker for SZ . The noticeable contribution of 1st person pronouns features we can observe in our models, including the singular form in the models that included 3rd level features, are in line with the work of Zomick et al [63] that also investigated Reddit data and found that increased 1st person singular pronouns are a indicator of SZ. Another conclusion our work shared with their findings was that decreased 1st person pronouns in the plural form was a characteristic of SZ. The pattern we found in the use of 1st person pronouns by SZ was reported as well in Strous et al. research [47] of written essays, which demonstrate that this characteristic appear in contexts different from social media communication. The anomaly in 1st person singular pronoun might reflect symptoms of SZ such as hyper-reflexivity related to loss of natural self-evidence and other disturbances of consciousness or self, as the 1st person plural

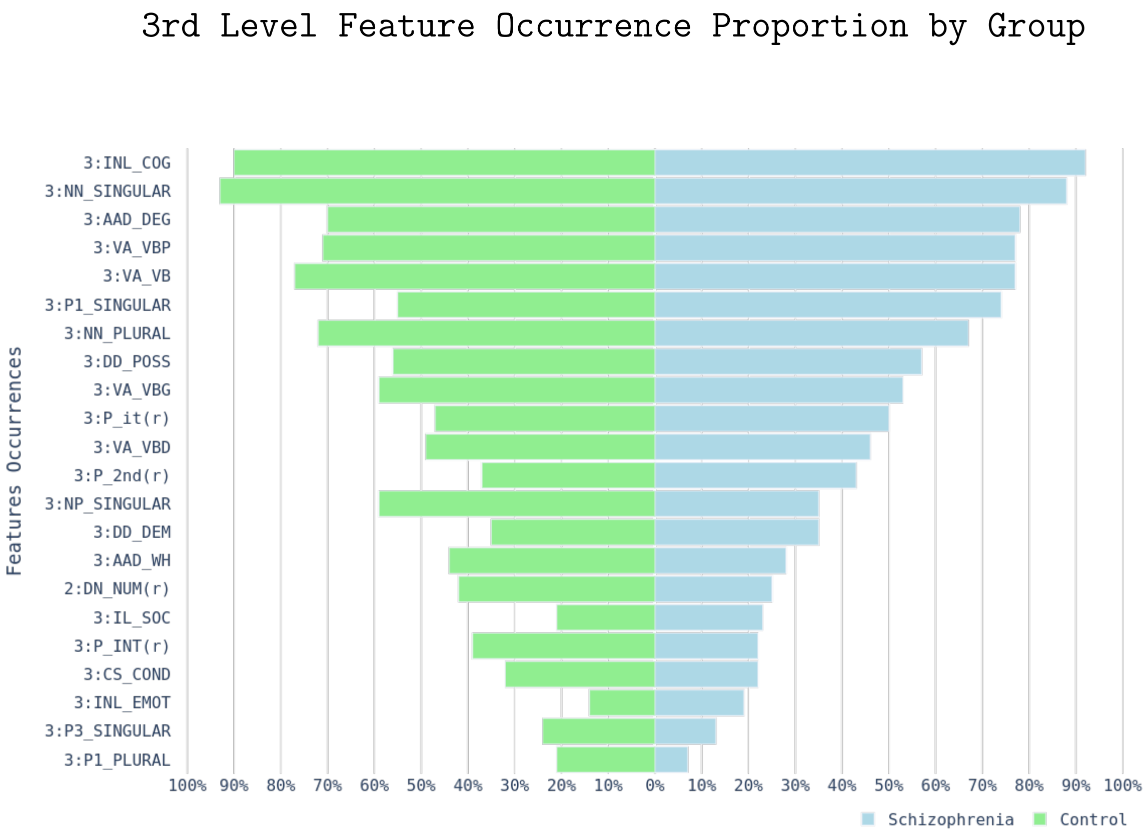


Figure 5.8: Percentage of feature occurrence 3rd Level Features for the SZ and Control Group

anomaly might be connected to other SZ traits as social disaffiliation and withdrawal.

One distinction of our results and the work of Zomick et al [63], was that our experiment pointed that both an abnormal increase or an abnormal decrease in 3rd person singular pronouns characterized SZ posts, while they concluded that only fewer 3rd person singular pronouns indicated SZ. Since the methods we used for classification and model explanation take into consideration the relationship among the features, this findings might reflect different patterns that appear in distinct sentence contexts. The overuse of 3rd person pronouns may reflect symptoms like externalizing bias and paranoid thinking, as in another context, the lack of 3rd person pronouns might represent problems with referentiality. The singular form of 3rd person pronouns emerged as a important feature in our work while the plural form was not as relevant.

Other features from the pronoun tree that stood out was interrogative pronouns, that SZ posts had fewer, and the pronoun *it*, witch SZ posts had more. Since pronouns encode person distinctions in grammar, problems in the use of pronouns, are possibly associated with difficulty in referentiality and definiteness [23] . Furthermore, when we consider that fewer proper nouns or fewer nouns in a broader sense, combined wit the overuse of the pronouns it are SZ markers, we might also consider this fact an indication of difficulty in the production of complex structures, which goes in the direction of research

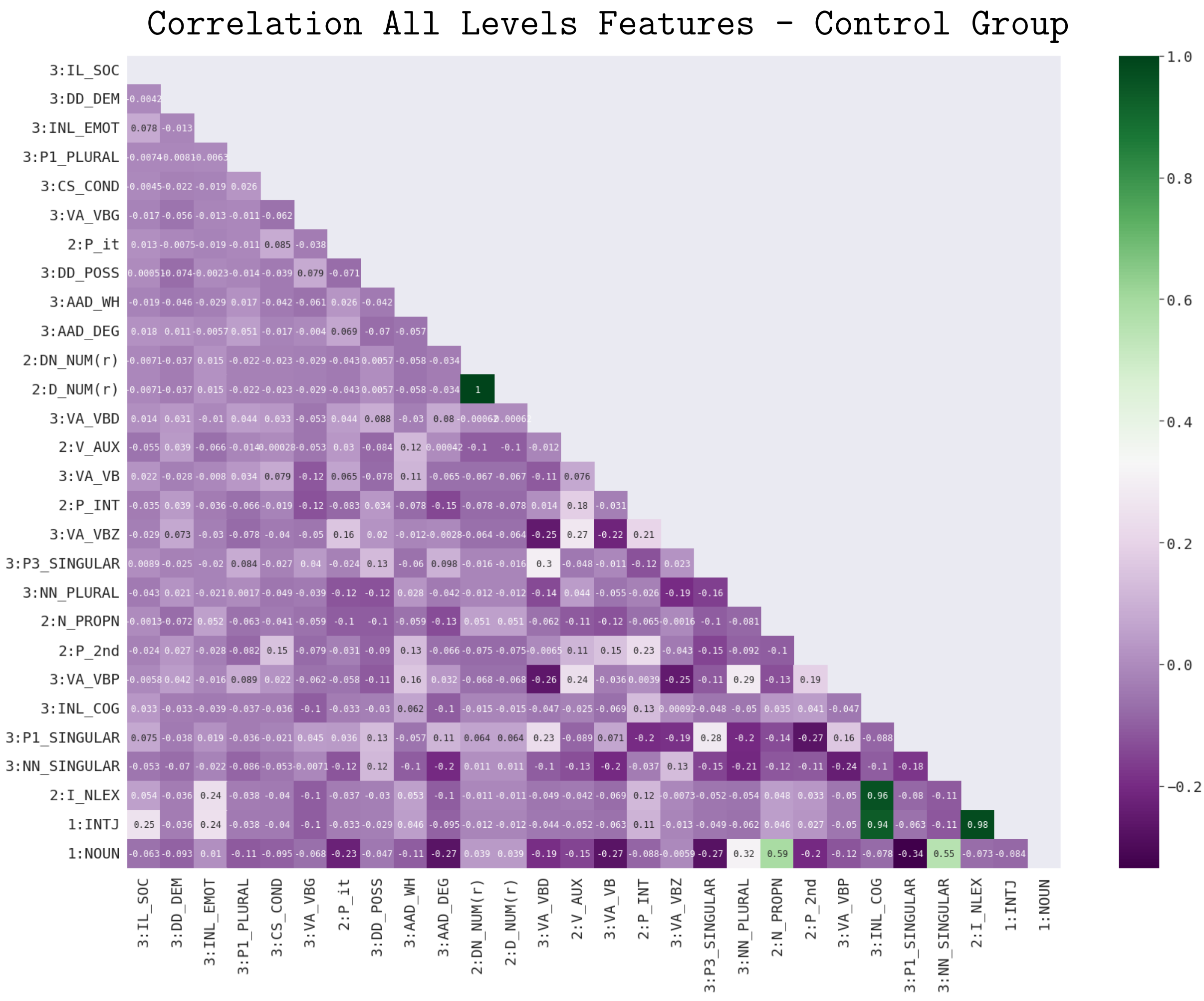


Figure 5.9: Pearson's Correlation Coefficient Matrices of Features from All Levels for the Control Group

in the synthetic level [10]. Overuse of the pronoun *it* deserves special attention since this is the 1st time this observation is made in respect of schizophrenia. The pronoun *it* is less specified semantically and statically than other pronouns, for example it does not contain gender, animacy or number information, which indicates that they have less syntactic structure. This hypothesis should be further examined, especially given that the correlation between the use of the it pronoun and schizophrenia symptoms was not established by previous works.

A feature from the noun tree that was selected by the 3rd level model and the model with the 3 levels, stood out for having divergent explanation. Higher value of common plural noun was used by the model with the tree levels as a marker for SZ, while the 3rd level model consider the opposite to be the case. Interestingly the two models are the ones with a higher agreement score. One possible explanation is that since the model with the tree levels had the opportunity of selecting the first level noun feature, and the second level proper noun feature, before selecting the third level common plural noun, it had a chance to establish a stronger pattern of influence of overall diminish presence of

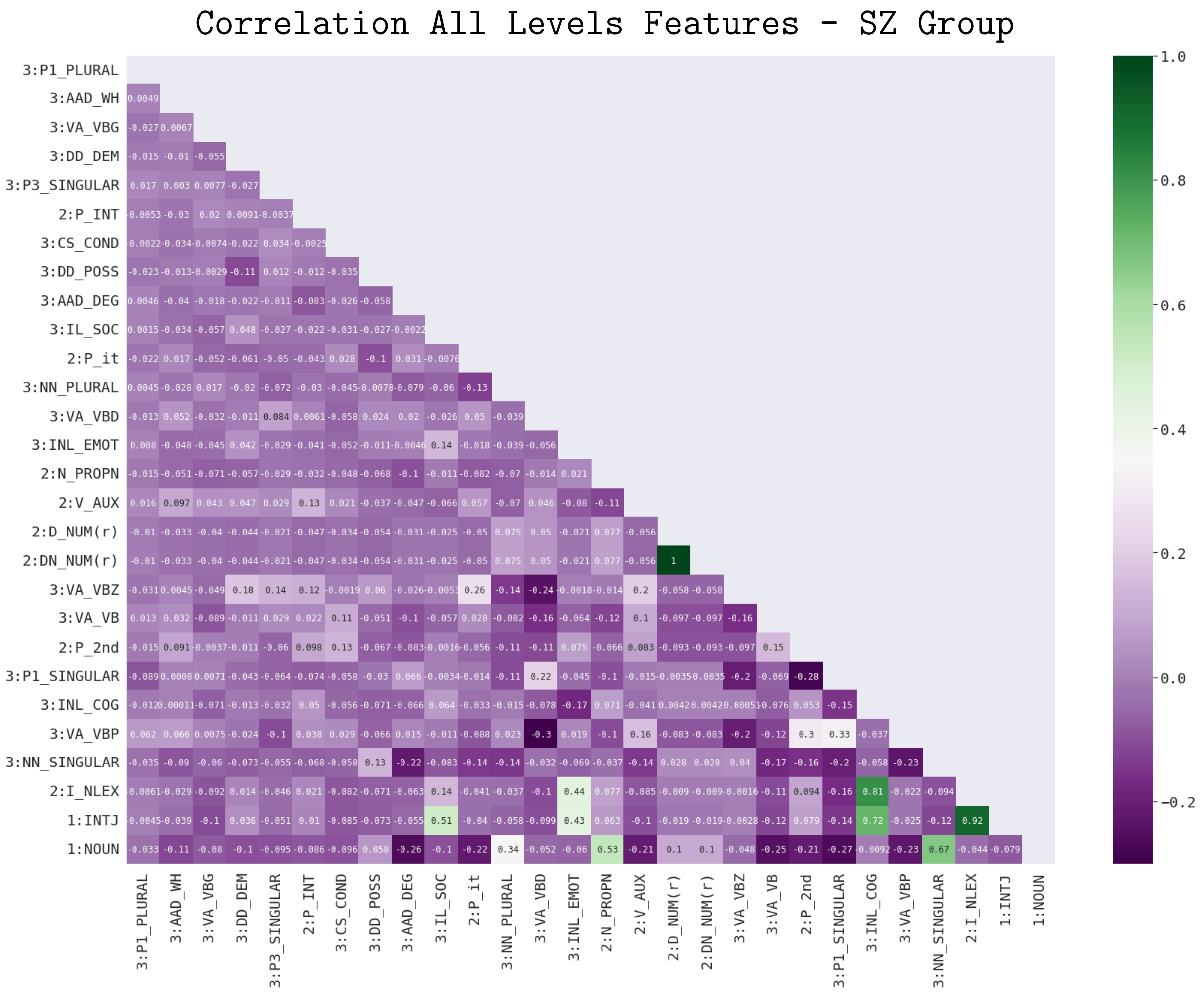


Figure 5.10: Pearson's Correlation Coefficient Matrices of Features from All Levels for the SZ Group

nouns, and further refine the specific role of plural common nouns. The third level model might have to combine the most representative noun tree features to determine the general impact of noun features, without the opportunity to single out characteristics that were only prominent in a certain group, or even needed the contrast with a more comprehensive feature to emerge. To be emphatic, this findings were only possible because we used a models with three levels of features specificity. Thus this findings illustrate gains in using such modeling methods. The overuse of common plural nouns might be related to a focus on outside groups and point to SZ symptoms of externalizing bias and paranoid thinking [16] [30] .

Similarly, auxiliary verbs and non lexical cognitive interjections had conflicting explanation. Since interjections was a high rank first level selected feature for the model with 3 levels, we can raise the same hypotheses we mention for the common plural nouns. As for auxiliary verbs a noteworthy combination of explanations aroused, with two third level auxiliary verbs features going on the opposite direction of the second level feature that preceded them, indicating that possibly the contrast among the features played an

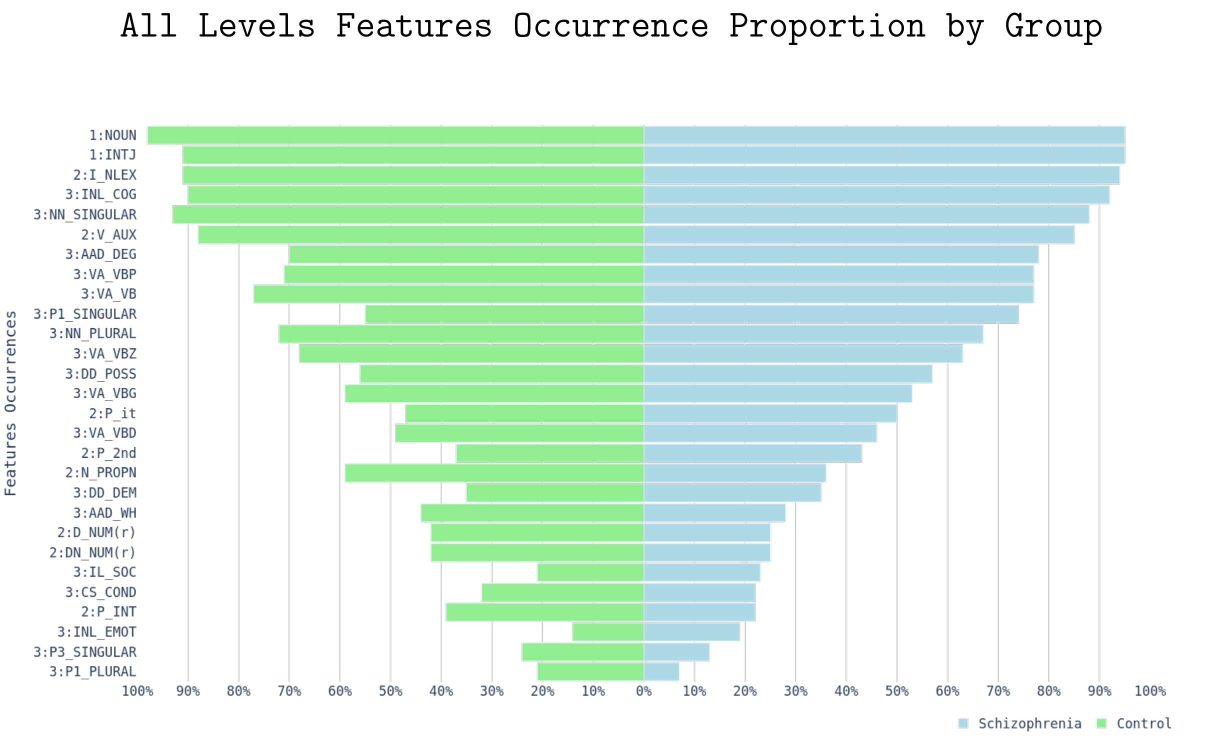


Figure 5.11: Percentage of feature occurrence from Features of All Levels for the SZ and Control Group

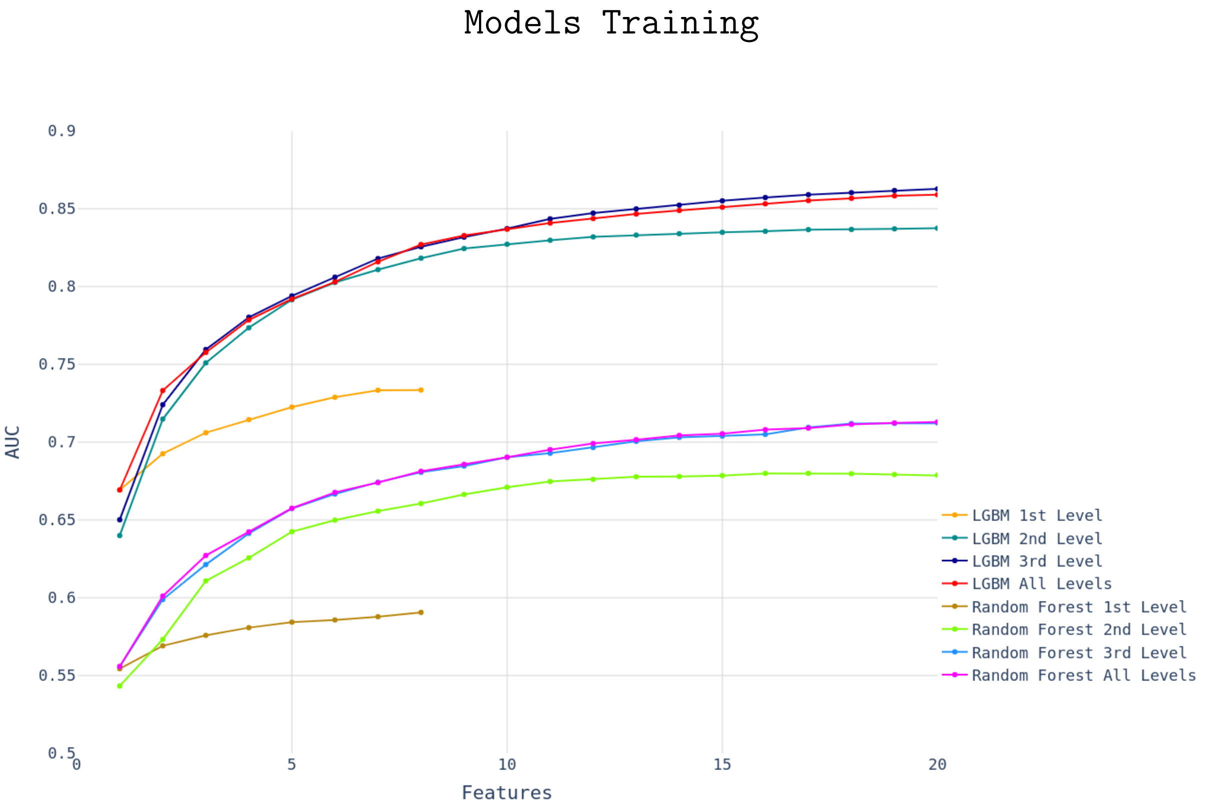


Figure 5.12: Training results: LGBM and Random Forest, all sets of features

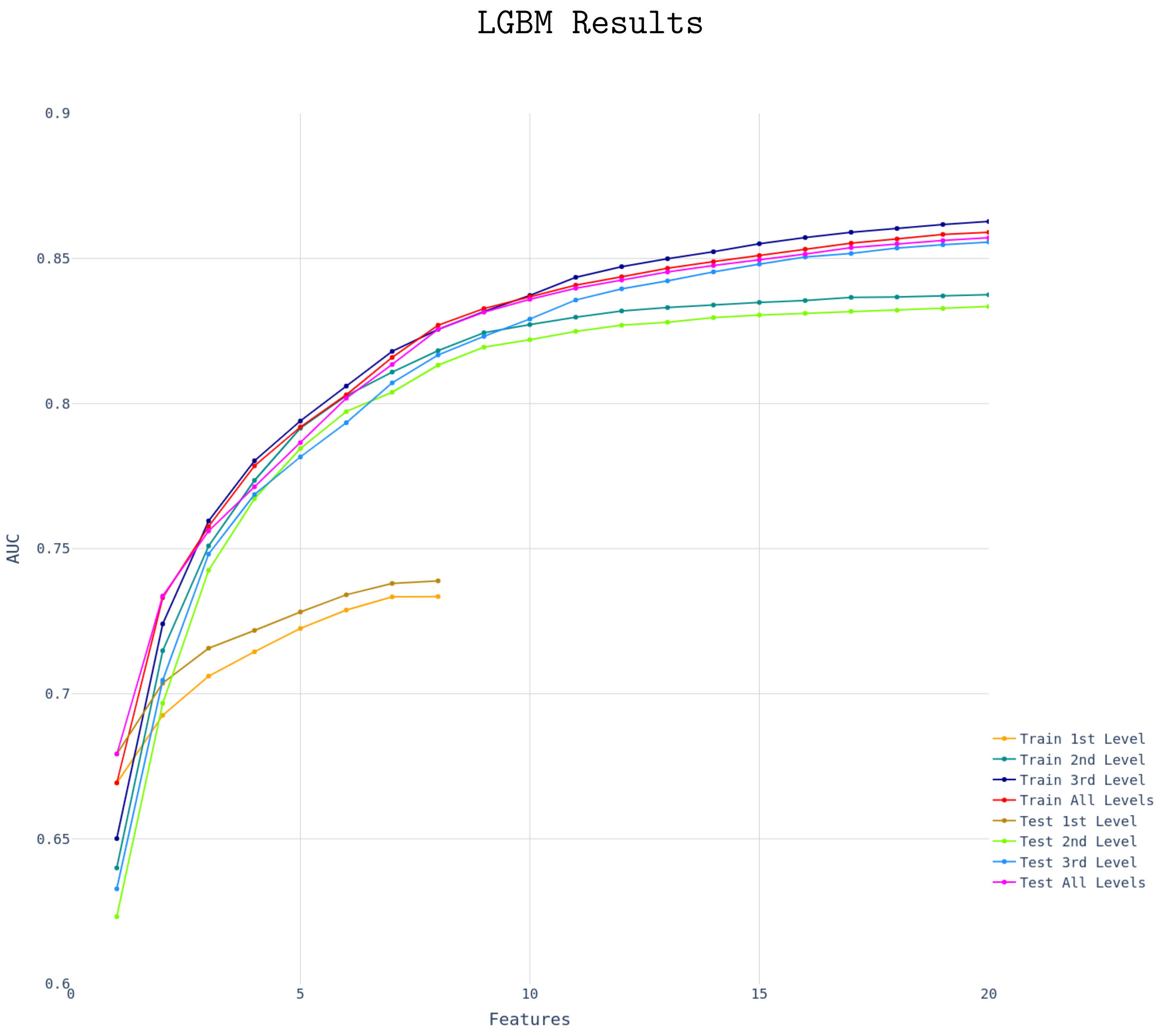


Figure 5.13: LightGBM training and test results

important role.

The distinction of interjection features noted in the rank of selected features was explained in the model with the 3 levels of features by more overall interjections and less non lexical cognitive interjections being markers for SZ. Interjections are highly related with organization of discourse in written language, and the presence of more anomalous lexical interjection might reflect poverty or disruption of grammatical structure, a reported characteristic of SZ expression. In the context of social media, that implies a more conventional informal aspect to the language, features related to non lexical interjection may also indicate communication incoherence as a hole and constant change in topic.

For illustration, on figures 5.17 and 5.18 we can observe local explanations of the two examples, of a post from the CT group and another for the SZ group, respectively. Bellow the plots, we can read the corresponding post texts. For the first example, we can see that all models have appropriately classified the CT post. The plots show that the three best models' top two most influential features were in the pronoun and noun trees.

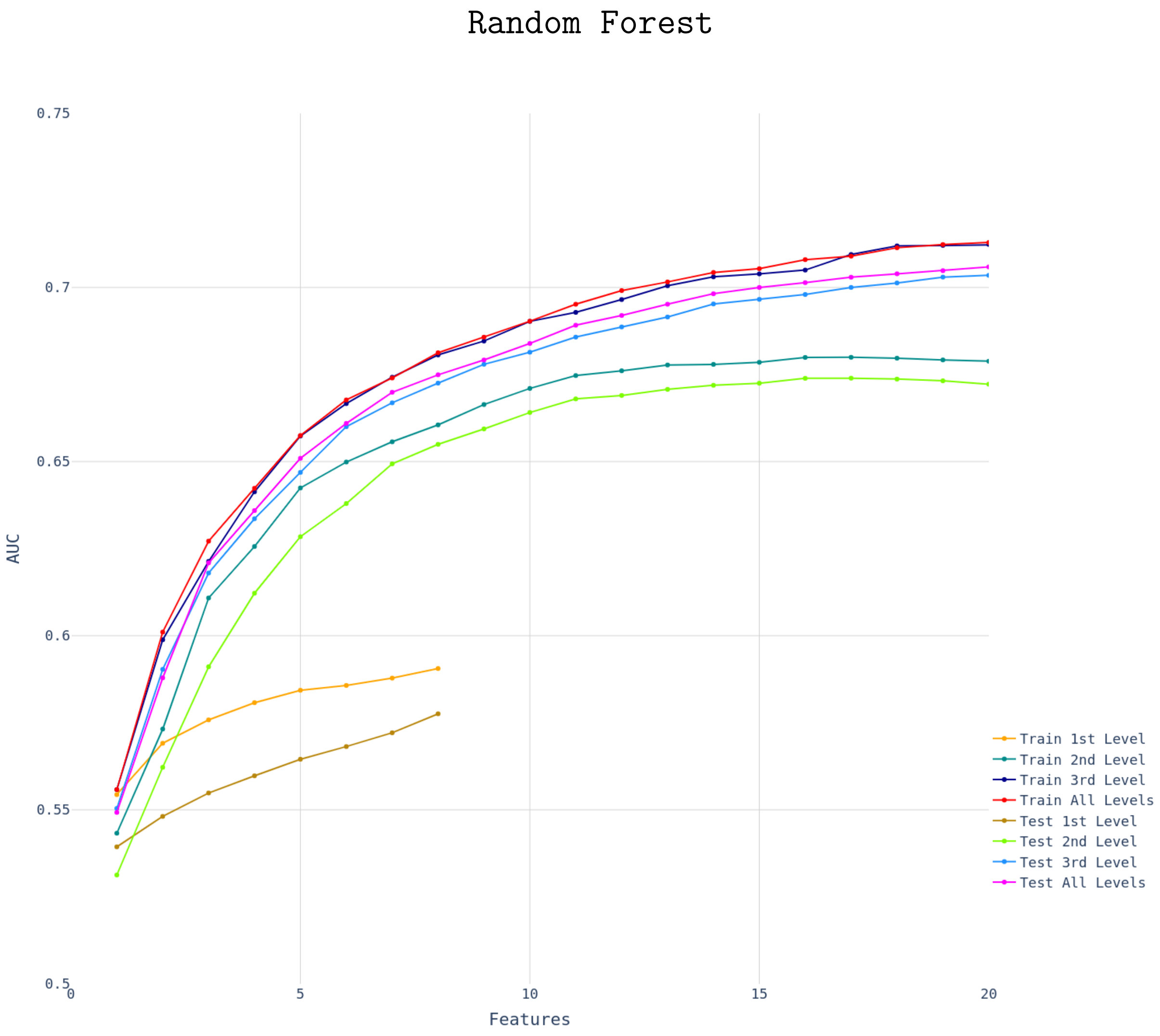


Figure 5.14: Random Forest training and test results

Features related to the use of first-person, singular forms, and the it pronoun, correctly impacted the result. On the other hand, features related the the use of second person and interrogative pronouns inaccurately pointed to the SZ group. Analyzing the post text, we can see that first-person pronouns are widely used, and is not surprising that the features that provided the most information were related to it.

Conversely, the second example illustrates an input that was wrongly classified by the first model and appropriately classified by the rest of the models. Similar to the first example, the text has a questioning nature and many first-person pronouns. Moreover, in this case, the plots also show that the top two most influential features of the tree best models were in the pronoun and noun tree, and the features that most correctly impacted the result were related to the use of first person and singular forms. Even though the texts shared characteristics and similar values for the most insightful features, except for the first-level model, the models were able to distinguish the correct class.

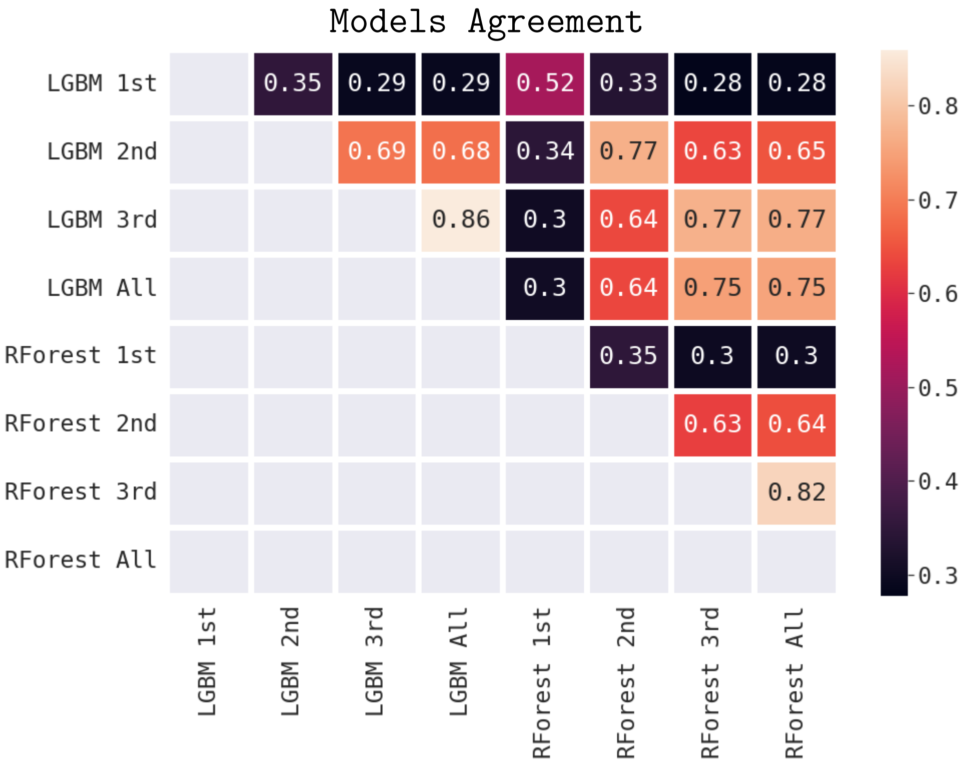


Figure 5.15: Models Kappa Coefficient Matrix

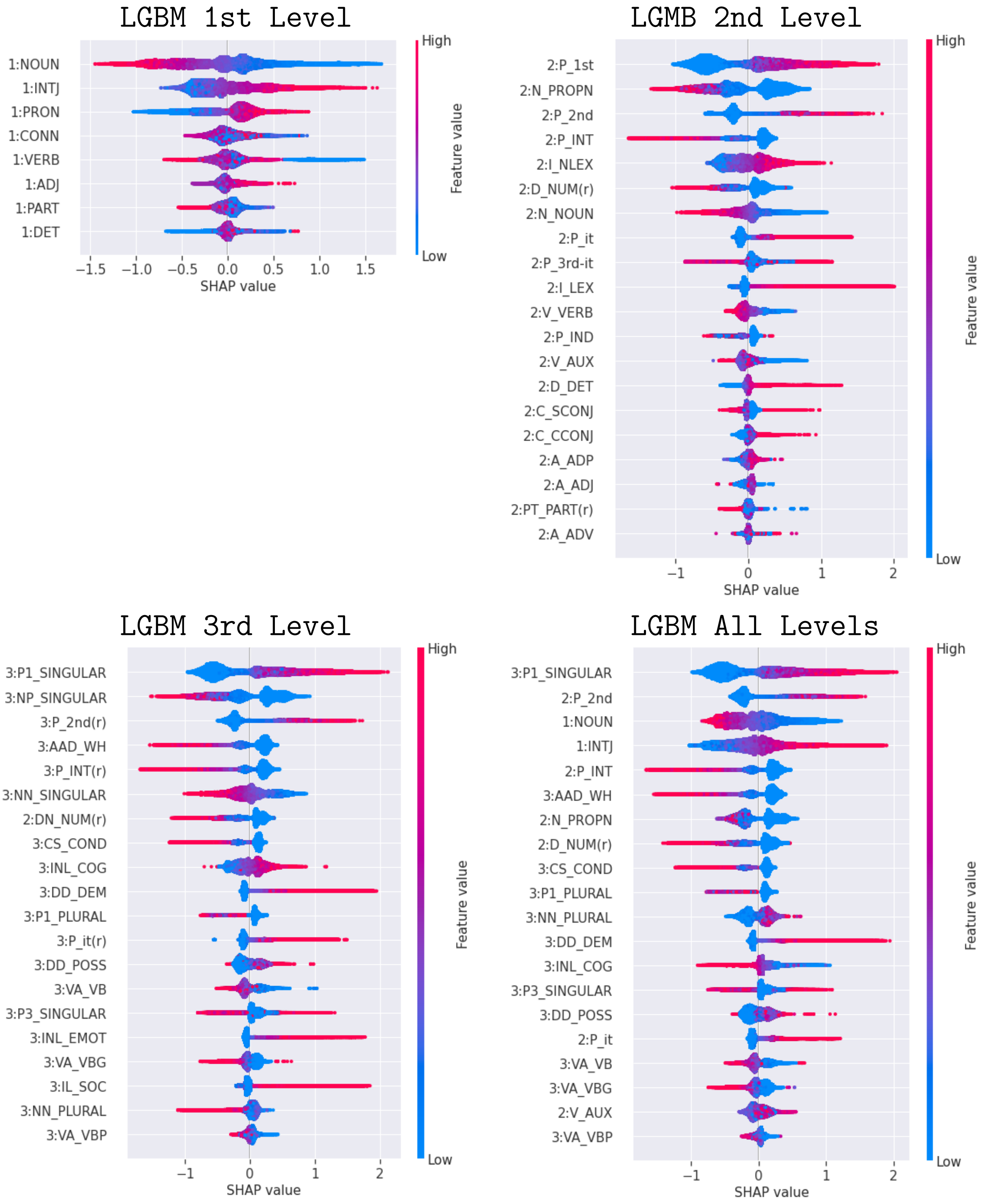
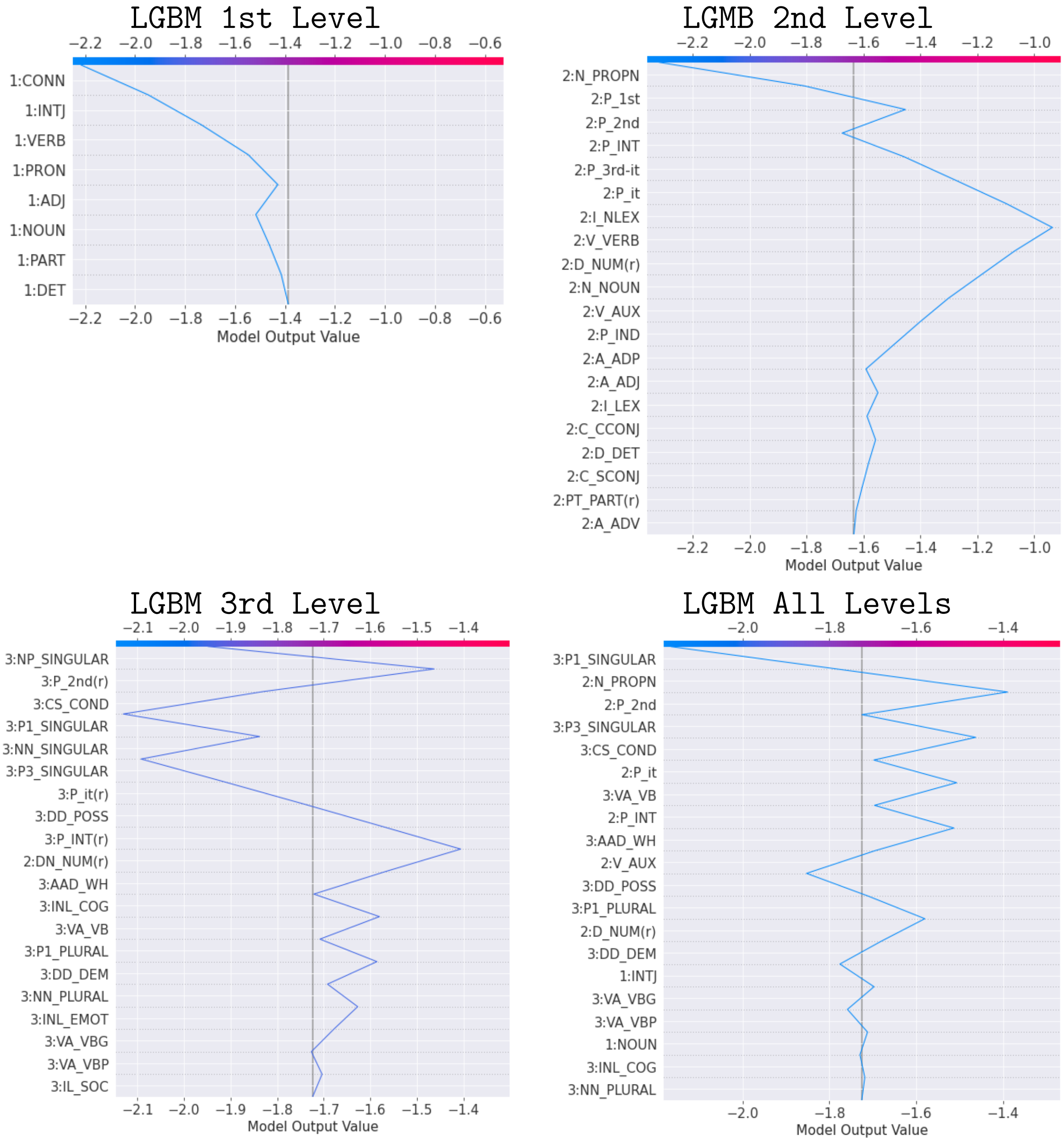
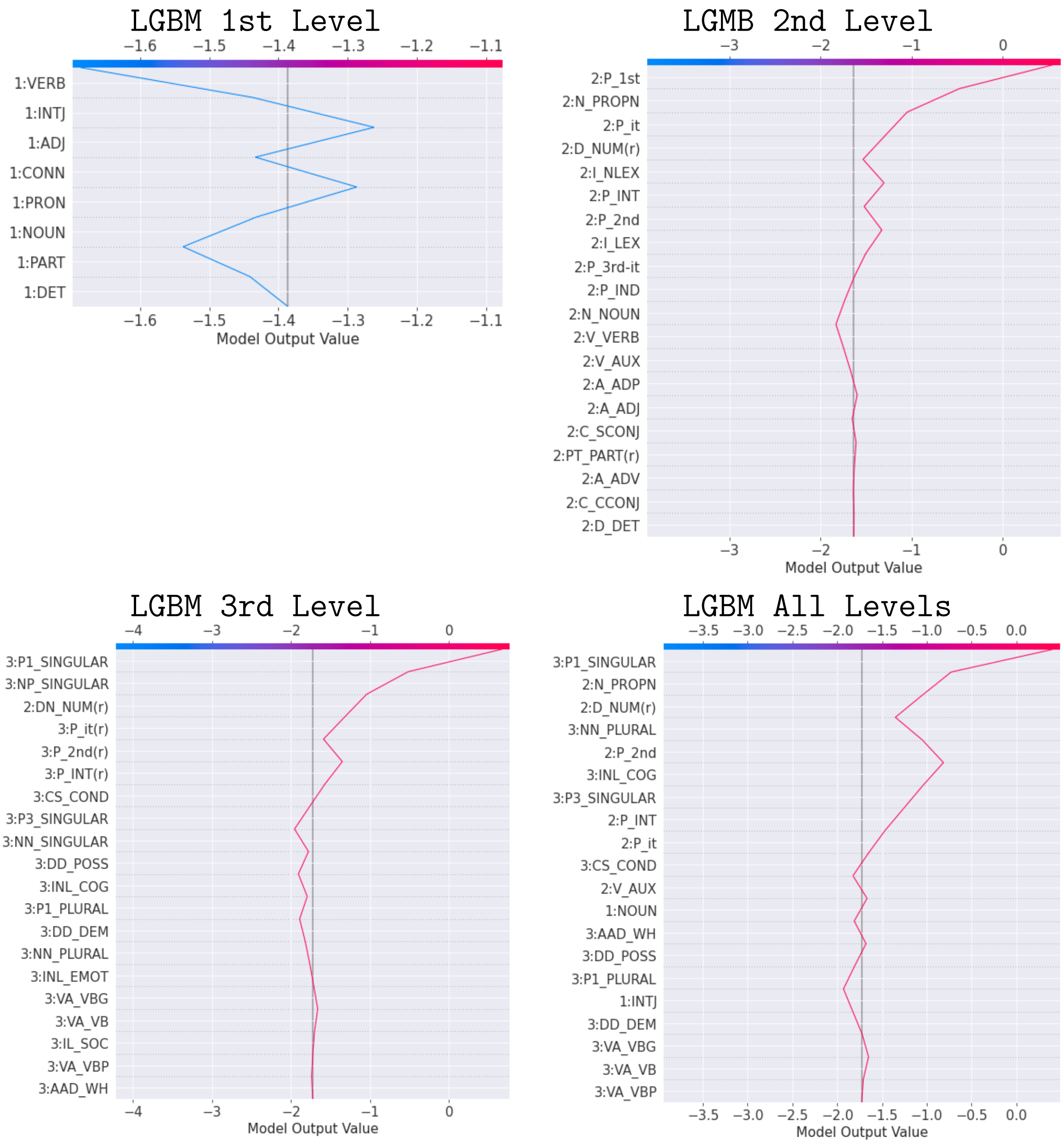


Figure 5.16: Shap Summary Plots of the best performing models of each set of fetures



Post: *'Could I sustain my water needs based solely off of salad ? So salad contains a lot of water, so if I was to eat a ton of salad each day, how much would I need to eat to get my water needs. Is it reasonable ? Doable ? Has anyone done it ? Note: It wouldn't be just eating salad, I could still have meat and such just can I eat a reasonableish salad amount to sustain my water needs. Note 2: Not actually thinking of doing this. I just like knowledge.'*

Figure 5.17: Shap decision plots for the first example of classification. the post is from the control group and was correctly classified by all the models.



Post: *'I need some advice for stomach tattoo.Hey I need some advice on stomach tattoo pain. I did my stomach a few days ago and well obviously it hurt like a bitch. But it was so bad I ended up getting sick as in I got a cold because I was probably to stressed and not relaxing good enough. Also fucking cold wind didn't help but anyhow. I really want the rest of my stomach tattooed and eventually that's gonna hit my ribs as well. How can I maybe make it easier for me the next time so I won't restrain my body so much ? '*

Figure 5.18: Shap decision plots for the second example of classification, the post is from SZ group and was correctly classified by three of the models

Chapter 6

Conclusions

The purpose of this study was to contribute to the understanding of the linguistic traits of SZ and advance the development of an exploratory methodology applicable to the study of language as a biomarker. To that end, we collected data from the social platform Reddit and created a specialized corpus for the classification of posts from people with SZ and CT. We designed three different hierarchical levels of POS tags using SpaCy and a rule-based system. We chose two algorithms, LightGBM and Random Forest, to train the models. Then we selected four sets of features for each algorithm, one from each level and a set combining the three levels. We trained and compared the model's results and explainability.

The LightGBM models of the third level and the combined levels were the best-performing ones. The model with combined levels of features showed less variance, with the smallest difference between training estimation and the test results, thus indicating that combining different levels of POS specificity is advantageous in this case. Since the train and test datasets were divided by the date of the beginning of the Covid-19 pandemic declaration, a low model estimation error also pointed to the fact that, even though symptoms might have worsened with the uncertainty and social isolation that characterized the test data period, they did not significantly change in nature or presentation.

Furthermore, we trained a transformer model to establish a baseline for the classification task and noted that the explanation it could provide was highly topic-dependent. Nevertheless, it helped in the assessment that POS features do allow for an assertive distinction between SZ and CT groups, living up to the data classification potential.

Analyzing results explanations, we were able to confirm previous studies' findings, namely, that more first-person singular pronouns, fewer first-person plural pronouns, and fewer interrogative pronouns are indications of SZ. We also observed a different pattern for third-person singular pronouns, noting that any discrepancy in this type of pronoun, either more or less than the mean, was a marker of SZ.

An important contribution of our work was the discovery of two novel SZ markers, the overuse of it pronouns and interjections, that in different levels of specificity, were identified among the most relevant features for classification. These observations might only have been possible with the employment of nonlinear models and in the context of written internet forums. More investigation is needed to understand these markers further and correlate them with cognitive symptoms.

A restriction of analyzing Reddit data was that SZ group was determined based on self-declaration with no evidence that they were clinically diagnosed, and there was no information on age, gender, first language, history of onset and symptoms, and similar data that could provide more experimental control and basis for more specific analysis and conclusions. Also, the distribution of the corpus data among SZ and CT was not representative of the general population, which would be of 1% SZ. Likewise, redditors of the Schizophrenia subreddit are subject to selection bias, in the sense that, they are not necessarily representative of their population.

Since our focus was on the methodological aspects of identifying general syntactic information, we did not further explore specific features. Future studies concentrating on how features can be grouped in distinct patterns might be able to provide further elucidations. Moreover, only linguistic features were used in the development of the model. Future research may take non-linguistic information into account, such as frequency and timing of the posts, changes in user activity level, and user online social involvement, incorporating an individual point of view to the analysis, verifying how individual makers' scores may vary with behavior and what might be the reason for that.

Bibliography

[1] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset, 2020.

[2] Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1(1):1–7, 2015.

[3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[4] Arthur A Berberian, Giovanna V Moraes, Ary Gadelha, Elisa Brietzke, Ana O Fonseca, Bruno S Scarpato, Marcella O Vicente, Alessandra G Seabra, Rodrigo A Bressan, and Acioly L Lacerda. Is semantic verbal fluency impairment explained by executive function deficits in schizophrenia? *Brazilian Journal of Psychiatry*, 38:121–126, 2016.

[5] Michael Leo Birnbaum, Sindhu Kiranmai Ernala, AF Rizvi, Elizabeth Arenare, A R Van Meter, M De Choudhury, and John M Kane. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from facebook. *NPJ schizophrenia*, 5(1):1–9, 2019.

[6] Nick Boettcher. Studies of depression and anxiety using reddit as a data source: Scoping review (preprint). *JMIR Mental Health*, 8, 04 2021.

[7] Benjamin Buck, Kyle S Minor, and Paul H Lysaker. Differential lexical correlates of social cognition and metacognition in schizophrenia; a study of spontaneously-generated life narratives. *Comprehensive psychiatry*, 58:138–145, 2015.

[8] Benjamin Buck and David L Penn. Lexical characteristics of emotional narratives in schizophrenia: relationships with symptoms, functioning, and social cognition. *The Journal of nervous and mental disease*, 203(9):702, 2015.

[9] Peter Calhoun, Xiaogang Su, Kelly Spoon, Richard Levine, and Juanjuan Fan. Random forest, 11 2021.

[10] Monica Chaves. *Structural Deficiency In Schizophrenia: An Exploratory Study Of The Nominal And Sentential Domains.* PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro, 05 2022.

[11] Linda A Clark and Daryl Pregibon. Tree-based models. In *Statistical models in S*, pages 377–419. Routledge, 2017.

[12] Mike Conway and Daniel O'Connor. Social media, big data, and mental health: current advances and ethical implications. *Current opinion in psychology*, 9:77–82, 2016.

[13] Nancy M Docherty. On identifying the processes underlying schizophrenic speech disorder. *Schizophrenia Bulletin*, 38(6):1327–1335, 2012.

[14] Helio Elkis. A evolução do conceito de esquizofrenia neste século. *Revista Brasileira de Psiquiatria*, 22, 05 2000.

[15] Brita Elvevåg, Peter W Foltz, Daniel R Weinberger, and Terry E Goldberg. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1-3):304–316, 2007.

[16] SK Fineberg, S Deutsch-Link, M Ichinose, T McGuinness, AJ Bessette, CK Chung, and PR Corlett. Word use in first-person accounts of schizophrenia. *The British Journal of Psychiatry*, 206(1):32–38, 2015.

[17] Institute for Health Metrics and Evaluation (IHME). Gbd compare, 2019.

[18] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE, 2020.

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning.* MIT press, 2016.

[20] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data? *arXiv preprint arXiv:2207.08815*, 2022.

[21] Tesfa Dejenie Habtewold, Lyan H. Rodijk, Edith J. Liemburg, Grigory Sidorenkov, Richard Bruggeman, and Behrooz Z. Alizadeh. Schizophrenia symptoms are inherently heterogeneous: a systematic review of cluster and group-based studies. *bioRxiv*, 2019.

[22] Heinz Häfner. From onset and prodromal stage to a life-long course of schizophrenia and its symptom dimensions: How sex, age, and other risk factors influence incidence and course of illness. *Psychiatry journal*, 2019, 2019.

[23] Wolfram Hinzen. Reference across pathologies: A new linguistic lens on disorders of thought. *Theoretical Linguistics*, 43, 09 2017.

[24] Wolfram Hinzen, Joana Rosselló, Cati Morey, Estela Camara, Clara Garcia-Gorro, Raymond Salvador, and Ruth de Diego-Balaguer. A systematic linguistic profile of spontaneous narrative speech in pre-symptomatic and early stage huntington's disease. *Cortex*, 100:71–83, 2018.

[25] Kai Hong, Ani Nenkova, Mary E March, Amber P Parker, Ragini Verma, and Christian G Kohler. Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls. *Psychiatry research*, 225(1-2):40–49, 2015.

[26] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[27] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[28] Gina R Kuperberg. Language in schizophrenia part 1: an introduction. *Language and linguistics compass*, 4(8):576–589, 2010.

[29] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[30] Minna Lyons, Nazli Deniz Aksayli, and Gayle Brewer. Mental distress and language use: Linguistic analysis of discussion forum posts. *Computers in Human Behavior*, 87:207–211, 2018.

[31] Linda McPherson. Discourse connectedness in manic and schizophrenic patients: Associations with derailment and other clinical thought disorders. *Cognitive neuropsychiatry*, 1(1):41–54, 1996.

[32] Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models using shapley values, 2019.

[33] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20, 2015.

[34] Tom Michael Mitchell. *The discipline of machine learning*, volume 9. Carnegie Mellon University, School of Computer Science, Machine Learning . . . , 2006.

[35] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[36] Natalia B Mota, Nivaldo AP Vasconcelos, Nathalia Lemos, Ana C Pieretti, Osame Kinouchi, Guillermo A Cecchi, Mauro Copelli, and Sidarta Ribeiro. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PloS one*, 7(4):e34928, 2012.

[37] World Health Organization. International programme on chemical safety biomarkers in risk assessment: Validity and validation, 2001.

[38] World Health Organization. Schizophrenia fact sheet, 2022.

[39] Lena Palaniyappan. More than a biomarker: could language be a biosocial marker of psychosis? *npj Schizophrenia*, 7(1):1–5, 2021.

[40] Marcia Radanovic, Rafael T de Sousa, L Valiengo, Wagner Farid Gattaz, and Orestes Vicente Forlenza. Formal thought disorder and language impairment in schizophrenia. *Arquivos de neuro-psiquiatria*, 71:55–60, 2013.

[41] Jason W Rocks and Pankaj Mehta. Memorizing without overfitting: Bias, variance, and interpolation in overparameterized models. *Physical Review Research*, 4(1):013201, 2022.

[42] Ribana Roscher, Bastian Bohn, Marco F Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *Ieee Access*, 8:42200–42216, 2020.

[43] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

[44] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2(3):1–21, 2021.

[45] Semrush. Most visited websites by traffic in the world, 2022.

[46] Leo Sher and René S Kahn. Suicide in schizophrenia: an educational overview. *Medicina*, 55(7):361, 2019.

[47] Rael Strous, Moshe Koppel, Jonathan Fine, Smadar Nachliel, Ginette Shaked, and Ari Zivotofsky. Automated characterization and identification of schizophrenia in writing. *The Journal of nervous and mental disease*, 197:585–8, 09 2009.

[48] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation, 2019.

[49] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[50] Rajiv Tandon, Wolfgang Gaebel, Deanna M Barch, Juan Bustillo, Raquel E Gur, Stephan Heckers, Dolores Malaspina, Michael J Owen, Susan Schultz, Ming Tsuang, Jim Van Os, and William Carpenter. Definition and description of schizophrenia in the dsm-5. *Schizophrenia research*, 150(1):3—10, October 2013.

[51] Debra Titone, Philip S Holzman, and Deborah L Levy. Idiom processing in schizophrenia: literal implausibility saves the day for idiom priming. *Journal of abnormal psychology*, 111(2):313, 2002.

[52] Antonia Tovar Torres, Wolfgang Sebastian Schmeisser Nieto, Aina Garí Soler, Catalina Morey Matamalas, and Wolfram Hinzen. Language disintegration under conditions of severe formal thought disorder. *Glossa: a journal of general linguistics*, 4(1), 2019.

[53] Danny et all Valdez. Social media insights into us mental health during the covid-19 pandemic: Longitudinal analysis of twitter data. *J Med Internet Res*, 22(12):e21418, Dec 2020.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[55] William N Venables and Brian D Ripley. Tree-based methods. In *Modern applied statistics with S-Plus*, pages 303–327. Springer, 1999.

[56] B. Venkatesh and J. Anuradha. A review of feature selection and its methods. *Cybernetics and Information Technologies*, 19(1):3–26, 2019.

[57] A. Voutilainen. Part-of-speech tagging. *The Oxford Handbook of Computational Linguistics*, 01 2012.

[58] Jeremy Watt, Reza Borhani, and Aggelos K Katsaggelos. *Machine learning refined: Foundations, algorithms, and applications*. Cambridge University Press, 2020.

[59] Daniel R Weinberger. Biological phenotypes and genetic research on schizophrenia. *World Psychiatry*, 1(1):2, 2002.

[60] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[61] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017.

[62] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

[63] Jonathan Zomick, Sarah Ita Levitan, and Mark Serper. Linguistic analysis of schizophrenia in reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 74–83, 2019.

Appendix A

Feature Description

The features were named according to the following rule: first a number indicating the feature's level, then a colon and the short form of the name of the preceding feature in the hierarchy, an underline symbol, and an abbreviation indicating the feature's main function. This rule was conceived to allow quick identification of the feature's tree, level and related part of speech. When a feature from the first or second level was not split into more subcategories, being just repeated in the next level, it was added the letter r between brackets to mark the repetition.

Feature	**Description**	**Level**	**Preceding**	**Examples**
1:PRON	pronouns	1st	-	me; you; itself; who
2:P_1st	1st person pronouns	2nd	1:PRON	me; mine, myself
3:P1_SINGULAR	singular 1st person pronouns	3rd	2:P_1st	i; myself; my
3:P1_PLURAL	plural 1st person pronouns	3rd	2:P_1st	we; us; ours; ourselves
2:P_2nd and 3:P_2nd(r)	2nd person pronouns	2nd and 3rd	1:PRON	you; u; yours; urself
2:P_3rd-it	3rd person pronouns and it	2nd	1:PRON	oneself; he; her; their
3:P3_SINGULAR	singular 3rd person pronouns	3rd	2:P_3rd-it	he; her; himself; one
3:P3_PLURAL	plural 3rd person pronouns and it	3rd	2:P_3rd-it	they; them; their
2:P_it and 3:P_it(r)	it pronouns	2nd and 3rd	1:PRON	it; its; itself
2:P_INT and 3:P_INT(r)	interrogative pronouns	2nd and 3rd	1:PRON	who; whom; what
2:P_IND	indicative pronouns	2nd	1:PRON	every; someone; none; all
3:PID_SINGULAR	singular indicative pronouns	3rd	2:P_IND	anything; something; one
3:PID_PLURAL	plural indicative pronouns	3rd	2:P_IND	everyone; everybody; ones
2:P_there and 3:PT_there(r)	there pronoun	2nd and 3rd	1:PRON	there

Table A.1: Description of the features from the Pronoun Tree

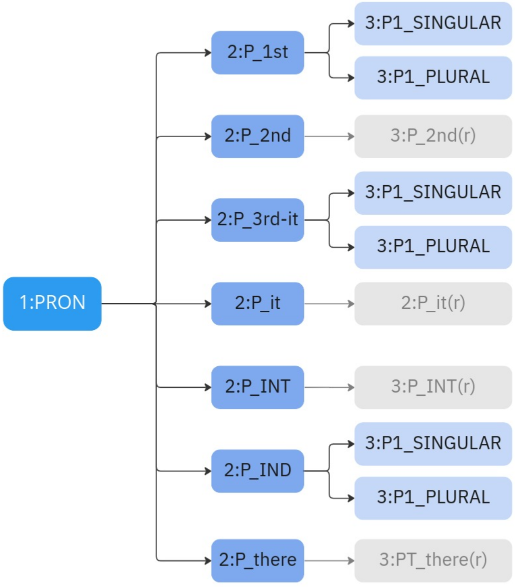


Figure A.1: Pronoun Tree visual representation

Feature	Description	Level	Preceding	Examples
1:NOUN	Nouns	1st	-	sofa; Rachel; bikes; street
2:N_NOUN	Common nouns	2nd	1:NOUN	house; medicine; places; voices
3:NN_PLURAL	Plural Common nouns	3dr	2:N_NOUN	books; drivers; cars; cats
3:NN_SINGULAR	Singular Common nouns	3dr	2:N_NOUN	boy; pencils; hospital; trees
2:N_PROPN	Proper nouns	2nd	1:NOUN	Anne; Europe; Joneses; Garfield
3:NP_PLURAL	Plural Proper nouns	3rd	2:N_NOUN	Smiths; Americas; Karens
3:NP_SINGULAR	Singular Proper nouns	3rd	2:N_NOUN	Mary; George; Australia

Table A.2: Description of the features from the Noun Tree

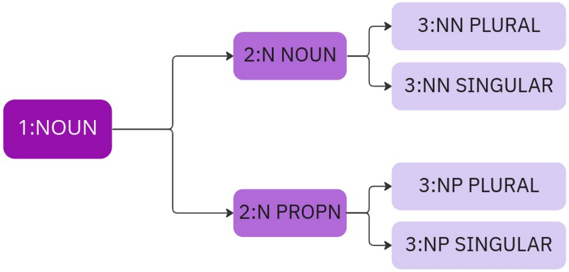


Figure A.2: Noun Tree visual representation

Feature	Description	Level	Preceding	Examples
1:CONN	Conjunctions	1st	-	that; plus; except; so; as
2:C_SCONJ	Subordinating Conjunctions	2nd	1:CONN	while; that; as; so; like
3:CS_TIME	Time Related Subordinating Conjunctions	3rd	2:C_SCONJ	while; when; soon; since; after
3:CS_that	'That' Subordinating Conjunctions	3rd	2:C_SCONJ	that
3:CS_COND	Conditional Subordinating Conjunctions	3rd	2:C_SCONJ	whether; if; unless
3:CS_COMP	Comparative Subordinating Conjunctions	3rd	2:C_SCONJ	than; like; as; rather
3:CS_ADD	Addition Subordinating Conjunctions	3rd	2:C_SCONJ	yet; and; plus; except
3:CS_is	'Is' Subordinating Conjunctions	3rd	2:C_SCONJ	is
3:CS_CASUAL	Casual Subordinating Conjunctions	3rd	2:C_SCONJ	because; so; due; since
3:CS_OTHERS	Other Subordinating Conjunctions	3rd	2:C_SCONJ	other subordinating conjunction
2:C_CCONJ	Coordinating Conjunctions	2nd	1:CONN	while; if; despite; so; nor
3:CC_TIME	Time Related Coordinating Conjunctions	3rd	2:C_CCONJ	while; when; once; after
3:CC_ADD	Addition Coordinating Conjunctions	3rd	2:C_CCONJ	yet; plus; despite; except
3:CC_CASUAL	Casual Coordinating Conjunctions	3rd	2:C_CCONJ	because, so; cause; providing
3:CC_COND	Conditional Coordinating Conjunctions	3rd	2:C_CCONJ	whether; if; unless; nor; either
3:CC_COMP	Comparative Coordinating Conjunctions	3rd	2:C_CCONJ	than, like, near, as, rather
3:CC_OTHER	Other Coordinating Conjunctions	3rd	2:C_CCONJ	other Coordinating conjunction

Table A.3: Description of the features from the Conjunction Tree

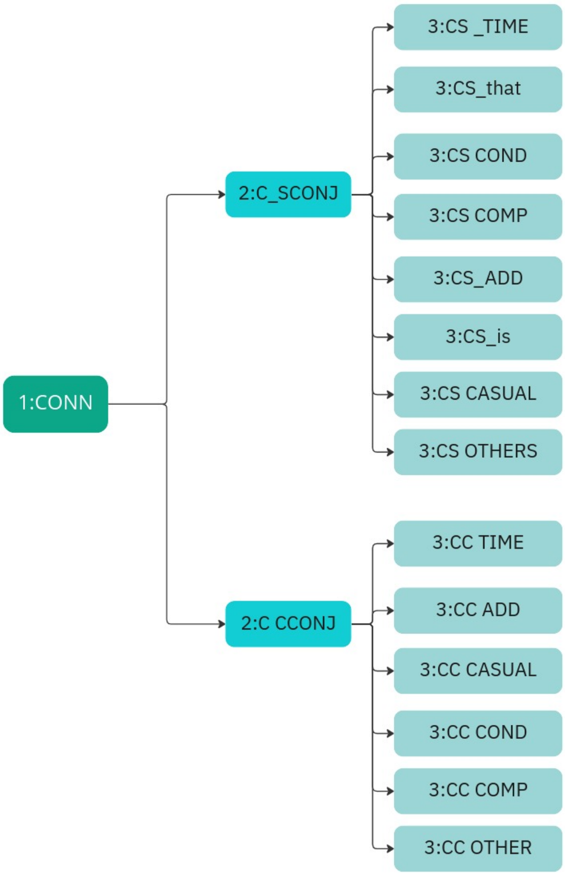


Figure A.3: Conjunction Tree visual representation

Feature	Description	Level	Preceding	Examples
1:DET	Determiners	1st	-	their; an; two; anyone; never; 5
2:D_DET	Non-Numeric Determiners	2nd	1:DET	these; anyone; his; once
3:DD_ART	Articles	3rd	2:D_DET	a; an; the
3:DD_DEM	Demonstrative	3rd	2:D_DET	this; that; those
3:DD_POSS	Possessive	3rd	2:D_DET	my; your; her; its
3:DD_QNT	Quantifier	3rd	2:D_DET	anything; someone; much
3:DD_ALT	Alternative	3rd	2:D_DET	others; another; which
3:DD_TIME	Time-related Determiners	3rd	2:D_DET	soon; once; whenever; due
3:DD_OTHERS	Other Determiners	3rd	2:D_DET	other determiners
2:D_NUM(r) and 3:DN_NUM(r)	numerals	2nd and 3rd	1:DET	one; 3; seventy

Table A.4: Description of the features from the Determiner Tree

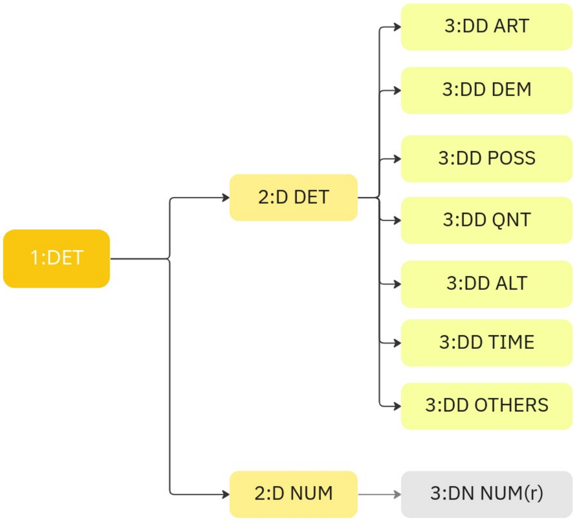


Figure A.4: Determiner Tree visual representation

Feature	**Description**	**Level**	**Preceding**	**Examples**
1:ADJ	Adjunct	1st	-	of; once; sooner; quite
2:A_ADP	Adposition	2nd	1:ADJ	to; from; of; under
3:AA_IN	”In” Adjunct	3dr	2:A_ADP	in
3:AA_RP	Particle	3dr	2:A_ADP	back; foward
2:A_ADV	Adverbs	2nd	1:ADJ	saddest; mostly; soon
3:AAD_COMP	Comparative Adverbs	3dr	2:A_ADV	faster; later; longer
3:AAD_SUP	Superlative Adverbs	3dr	2:A_ADV	soonest; loudest; quickest
3:AAD_DEG	Degree Adverbs	3dr	2:A_ADV	almost; barely; entirely; highly
3:AAD_WH	Wh-Adverbs	3dr	2:A_ADV	when; where; why
2:A_ADJ	Adjectives	2nd	1:ADJ	their; who; such; -ing
3:AADJ_AFX	Affixes	3rd	2:A_ADJ	un-; self-; pre-; re-
3:AADJ_PDT	Predeterminer	3rd	2:A_ADJ	twice; such; quite; half; both
3:AADJ_POSS	Possessive	3rd	2:A_ADJ	my; her; our
3:AADJ_WDT	Wh-determiner	3rd	2:A_ADJ	which; what; who
3:AADJ_WPR	Wh-pronoun	3rd	2:A_ADJ	whose; who; whether

Table A.5: Description of the features from the Adjunct Tree

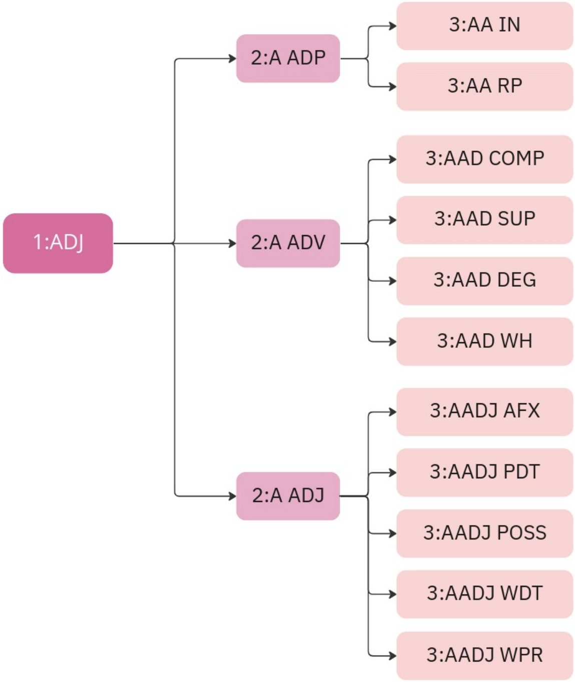


Figure A.5: Adjunct Tree visual representation

Feature	Description	Level	Preceding	Examples
1:PART and 2:PT_PART(r)	Particles	1st and 2nd	-	-n't; -'s; of
3:PT_POSS	Possessive Ending	3rd	2:PT_PART(r)	-'s; -s'
3:PT_RP	Adverb Particle	3rd	2:PT_PART(r)	back; down; of
3:PT_TO	Infinitival	3rd	2:PT_PART(r)	to
3:PT_RB	Negative	3rd	2:PT_PART(r)	-n't; not; -un

Table A.6: Description of the features from the Particle Tree

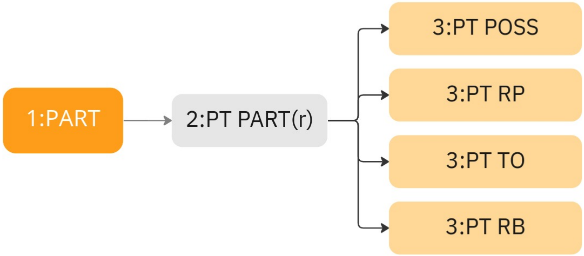


Figure A.6: Particle Tree visual representation

Feature	Description	Level	Preceding	Examples
1:INTJ	Interjections	1st	-	ho, !; wow; ugh; ?
2:I_LEX	Lexical Interjections	2nd	1:INTJ	hey; bravo; ahh; ish
3:IL_EMT	Emotive Lexical Interjections	3rd	2:I_LEX	nooo; ohh; ugh; yippee
3:IL_SOC	Social Lexical Interjections	3rd	2:I_LEX	Wow!; dang; shoo; shh
2:I_NLEX	Non-Lexical Interjections	2nd	1:INTJ	.; !; ?; -;
3:INL_EMOT	Emotional Non-Lexical Interjections	3rd	2:I_NLEX	...; !
3:INL_COG	Cognitive Non-Lexical Interjections	3rd	2:I_NLEX	;; ?; -

Table A.7: Description of the features from the Interjection Tree

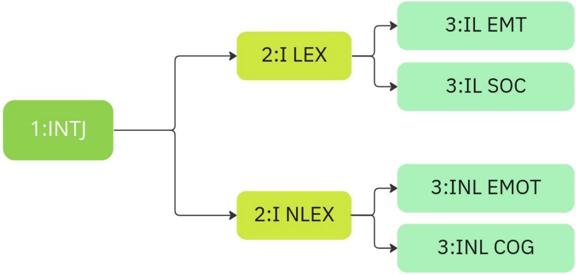


Figure A.7: Interjection Tree visual representation

Feature	Description	Level	Preceding	Examples
1:VERB	Verbs	1st	-	doing; need; was; could
2:V_AUX	Auxiliary Verbs	2nd	1:VERB	were; should; does; go
3:VA_VBD	Past Tense Auxiliary Verbs	3rd	2:V_AUX	was; were
3:VA_VBP	Non-3rd Person Singular Present	3rd	2:V_AUX	want; need
3:VA_VBG	Present Participle	3rd	2:V_AUX	going; doing
3:VA_VBZ	3rd Person Singular Present Verbs	3rd	2:V_AUX	does; leaves; goes
3:VA_MD	Modal Auxiliary Verbs	3rd	2:V_AUX	could; should
2:V_VERB	Non-Auxiliary Verbs	2nd	1:VERB	eat; wants; went; find
3:VB_VBN	Past Participle Verb	3rd	2:V_VERB	lost; found; looked
3:VB_VBG	Gerund	3rd	2:V_VERB	looking; cooking; talking
3:VB_VB	Verb Base Form	3rd	2:V_VERB	go; stay; eat; fly
3:VB_VBZ	3rd Person Singular Present Verbs	3rd	2:V_VERB	wants; jumps; goes
3:VB_VBD	Past Tense Verbs	3rd	2:V_VERB	went; looked; left; said

Table A.8: Description of the features from the Verb Tree

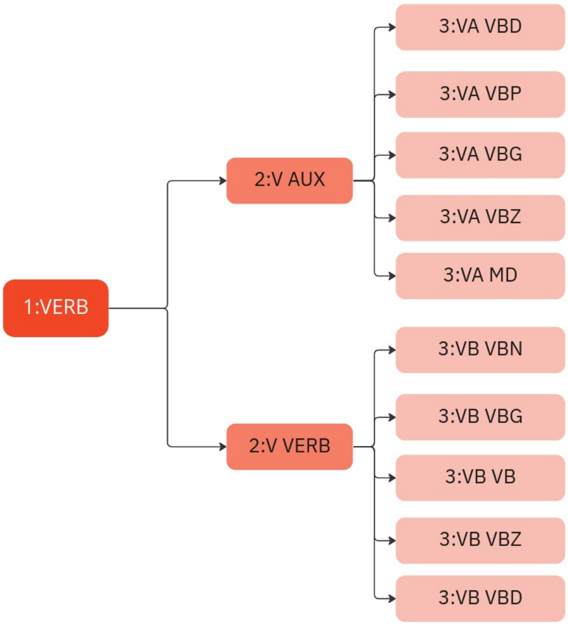


Figure A.8: Verb Tree visual representation