IberSPEECH 2022
14-16 November 2022, Granada, Spain

Extraction of structural and semantic features for the identification of Psychosis in European Portuguese

Rodrigo Sousa[1], *H. Sofia Pinto*[1], *Alberto Abad*[1], *Daniel Neto*[2], *Joaquim Gago*[3]

[1]INESC-ID Lisboa, Portugal
[2]Serviço de Saúde da Região Autónoma da Madeira, Portugal
[3]Nova Medical School/Centro Hospitalar de Lisboa Ocidental, Lisboa, Portugal

rbps@ssh.hlt.inesc-id.pt, sofia@inesc-id.pt, alberto.abad@inesc-id.pt, danielcarvalhoneto@gmail.com, joaquimgago@me.com

Abstract

Psychosis is a brain condition that affects the subject and the way it perceives the world around, impairing its cognitive and speech capabilities, and creating a disconnection from reality in which the subject is inserted. Psychosis lacks formal and precise diagnostic tools, relying on self-reports from patients, their families, and specialized clinicians. Previous studies have focused on the identification and prediction of psychosis through surface-level analysis of diagnosed patients targeting audio, time, and paucity features to predict or identify psychosis. More recent studies have started focusing on high-level and complex language analysis such as semantics, structure, and pragmatics. Only a reduced number of studies have targeted the Portuguese language. Currently, no study has targeted structural or semantic features in European Portuguese, thus this is our objective. The results obtained through our work suggest that the use of structural and semantic features, particularly for European Portuguese, holds some power in classifying subjects as diagnosed with psychosis or not. However, further research is required to identify possible improvements to the techniques employed and to concretely identify which particular features hold the most power during the classification tasks.

Index Terms: psychosis, schizophrenia, coherence analysis, structure analysis, content analysis, natural language processing, classification

1. Introduction

In 2017, it was reported that at the time around twenty million people suffered from psychosis, which represents 0.3% of the world population [1]. Specifically, regarding Portugal, a recent study reported that around 3% to 4% of the Portuguese population has at one point suffered from psychotic disorders [2].

According to the World Health Organization [3], "psychoses, including schizophrenia, are characterized by distortions in thinking, perception, emotions, language, sense of self, and behavior. Common psychotic experiences include hallucinations (hearing, seeing, or feeling things that are not there) and delusions (fixed false beliefs or suspicions that are firmly held even when there is evidence to the contrary)".

1.1. Motivation

Psychosis is marked by various symptoms, such as hallucinations, typically involving sounds or voices, sudden reclusiveness, longer pauses' duration, frequent and unnatural repetitions, disorganized or completely incoherent speech, and poor speech, marked by short sentences with low complexity. Due to inherent aspects of the disorder and the stigma propagated, patients who suffer from psychosis feel isolated from society.

The successful reintegration and recovery of diagnosed patients can be improved by early detection through preceding indicators [4], which justifies possible changes to diagnostic tools in terms of efficiency. Computerized aided diagnoses have long been introduced into other domains. Even though studies have proposed solutions in the domain of mental disorders and specifically for psychosis, none have been employed in real diagnosis. Such solutions have distinguished control from diagnosed patients [5, 6, 7, 8, 9, 10, 11] and even predicted future psychosis [12, 13, 14]. Computerized solutions mainly help diagnosis and prevention: for their efficiency, lower requirements in specialized training needed for their use, and ability to detect humanly imperceptible speech deviations.

Moreover, previous research work either differentiates controls from diagnosed patients [5, 6, 7, 12, 8, 9, 13, 14, 10, 11], patients with varying degrees of the disorder, [15, 16, 17, 18] or sometimes, although rarely, patients with different disorders [19, 20, 21]. When using computerized solutions, it is important that they also consider the various disorders, distinguishing them, and that such solutions are studied worldwide.

Lastly, to the best of our knowledge, only one study has focused on the European Portuguese language, the work of Forjó et al. [6]. The authors collected the first European Portuguese corpus of speech, from patients diagnosed with psychosis and controls. The mentioned corpus is composed of recordings from patients, following a protocol that does not require private or medical information. [6] developed a model capable of distinguishing patients diagnosed with psychosis from healthy controls through speech fluency and acoustic features.

1.2. Objectives

Our work took as a starting point the work of Forjó et al. [6]. We extended the previous work by exploring discourse's coherence, semantics, and content. Then, we evaluated whether the results could be supported or improved with these features. Additionally, we extended the corpus developed by [6], following the same protocol, with recordings from patients diagnosed with psychosis and subjects who are diagnosed with other mental disorders, specifically, bipolar disorder in its various stages. By expanding the current European Portuguese corpus for psychosis we can assess whether the results obtained for the corpus by [6] reflect inherent differences between psychotic subjects and the rest of the population or the classifier adjusted to other factors. [6] does not reflect the diversity that exists in the world, since, in the study, a subject is either healthy, showing no other disorders or signs thereof, or is diagnosed patient with

10.21437/IberSPEECH.2022-21

psychosis. Therefore, it is possible that the developed classifier targeted other factors, possibly, side-effects of the medication, which have been shown to affect and be detectable [22].

Our work does not attempt to distinguish the various mental disorders, but instead on identifying which subjects are diagnosed with psychosis in a population of healthy controls, patients diagnosed with psychosis, and patients diagnosed with other mental disorders. In Portugal, patients usually only contact clinicians after the first episode of psychosis. Consequently, our work aims only at identifying patients already diagnosed with psychosis and not at predicting psychosis.

2. Methods

Several methods were used in our work that need to be defined. The methodologies employed are separated into methodologies for corpus acquisition and treatment, methodologies for the extraction of fluency and acoustic features, and methodologies for the extraction of coherence, structural and content features, in subsections 2.1, 2.2, and 2.3 respectively.

2.1. Corpus Acquisition and Treatment

Our study relies on the extension of the corpus in [6], Figure 1. The original version, collected data from 92 subjects, where 56 were healthy subjects and 36 subjects were diagnosed with psychosis. During our research, we extended the existing corpus with 85 subjects, where 9 were healthy, 47 were diagnosed with psychosis, and 29 were diagnosed with bipolar disorder. Note that, per the ethics committees' agreement, the corpus developed will not be available to the public.

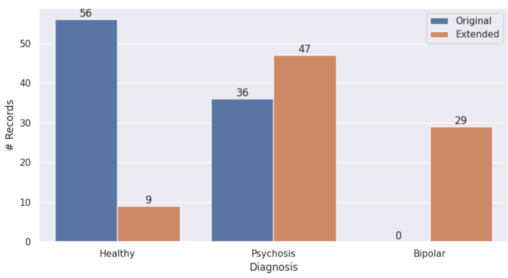

Figure 1: *Corpus used for research (original and extension).*

2.1.1. Protocol

Since our work aimed at extending the corpus already developed, we followed the same protocol as Forjó et al. [6]. The mentioned protocol takes approximately 20 minutes organized in seven tasks: (1) phonetic verbal fluency task in which the subjects enumerate words starting with 'p' for 60 seconds; (2) categorical verbal fluency task in which the subjects enumerate animals for 60 seconds; (3) reading of a well-known children's story; (4) retelling of a well-known children's story; (5) description of a positive affective image; (6) description of a neutral affective image; (7) description of a negative affective image. We also registered demographic information regarding the subjects, namely sex, age, education level, and particularly for the diagnosed patients, the duration of the diagnosis, to better characterize the corpus and compare groups' characteristics.

2.1.2. Recordings Processing

We used the previous transcription of the various recordings to assess speech coherence, structure, and content. Our team decided on using Tribus, an already developed transcriber for European Portuguese, developed by Carvalho and Abad [23]. We chose Tribus since it had already been used in [6], it is efficient, free, and works offline without requiring the submission of the audio files to online tools, possibly violating the patient's consent. It is noteworthy to mention that Tribus does not segment the transcribed text into sentences and that no good alternative exists for sentence segmentation of European Portuguese. Therefore, in the following methods defined hereafter, when segmentation of speech into sentences is required, we chose to segment the transcribed text into sets of words of equal length, basing ourselves on the technique of n-grams [24, 25].

An important step for various of the methods defined hereafter rely on word mappings, either from word embedding models or lexicons. Thus it is imperative to reduce the number of word inflections present in the transcriptions. We experimented with various lemmatizers and decided on using *Stanza* [26], due to its simplicity, efficiency, and results.

2.2. Speech Fluency and Acoustic analysis

Since in our work, we aim to compare the results obtained through speech fluency and acoustic features with the results that can be obtained with coherence, structural and content features, we developed the solution described in [6]. However, it is still meaningful to mention, that this model was not the focus of our work and served merely to achieve baseline values for the results obtained.

We extracted as speech features: (1) the number of words spoken during the recording; (2) the number of syllables spoken during the recording; (3) the duration of recording, in seconds; (4) the speaking rate of the subject during the recording, in words per second; and (5) the articulation rate of the subject during the recording, in syllables per second.

The acoustic features were computed through *GeMAPS* [27, 28], a software that allows for the automatic extraction of speech features from audio recordings. In our model, we used *eGeMAPS*, an expansion of *GeMAPS*, used by [6], which outputs 88 features instead of 62 features.

2.3. Coherence, Structural and Content Analysis

As mentioned, the focus of our work was on the features extracted with the methods that rely on the speech's content, structure, and coherence.

2.3.1. Latent Semantic Analysis' Features

Latent Semantic Analysis (LSA) consists of an extrapolation of the latent meaning of a word or passage, developing an embedding for it. Word embeddings are developed, from a corpus, through analysis of co-occurrences of these words with other words in a set of documents. It is important to mention that LSA computes embeddings independent of context. Therefore these embeddings are unable to correctly capture the various meanings that can be given to a word according to its context.

For this purpose, we used CETEMPúblico [29], a corpus made up of 1.485.828 extracts of articles published in a European Portuguese journal to achieve the mentioned word embeddings. One important step of LSA is the selection of the dimensionality to which the word embeddings are reduced. For this reason, we studied the coherence achieved for various levels of dimensionality reduction and attained the best results for a dimensionality of 20.

Finally, we segmented the text into groups of words, averaged the word embeddings of each group, and followed the methods of [7, 12, 16], computing the first and second-order coherence. The first-order coherence is calculated by computing the average of the cosine difference between each group and the subsequent group. The second-order coherence is calculated by computing the average of the cosine difference of each group and the group two positions ahead. In literature, patients diagnosed with psychosis, are identified as having low coherence [12, 16, 8, 17], marked by sudden changes in discourse topic.

2.3.2. Latent Content Analysis' Features

Following the methodology of Rezaii et al. [14], we start by selecting the 95% most common words, in our case, for European Portuguese, using once again the *CETEMPúblico* [29] corpus. The corpus served not only to identify the most common words but to compute *Word2Vec* [30] embeddings for each word, which were averaged out, according to groups.

Then, for each transcription and each one of the most frequent words, we compute the highest cosine between any of the transcription groups and the frequent word. Subsequently, we selected the top 50 words that maximize the difference of cosines between groups as well taking into account their prevalence amongst all the documents by applying *TF-IDF*. Effectively we achieved the 50 words most prevalent in meaning in one group and least in the other. From these words we developed clusters, choosing the number of clusters that maximize their silhouette coefficient. An example of the clusters developed can be seen in Figure 2.

As features for the classification task, we required a value that expressed the similarity of the transcription with each one of the clusters. To this objective, we computed the highest cosine value from the various group embeddings, of a given transcription, with each one of the cluster centers computed. We expected these clusters to differ according to the group, and the distances to these clusters to express this difference in topic.

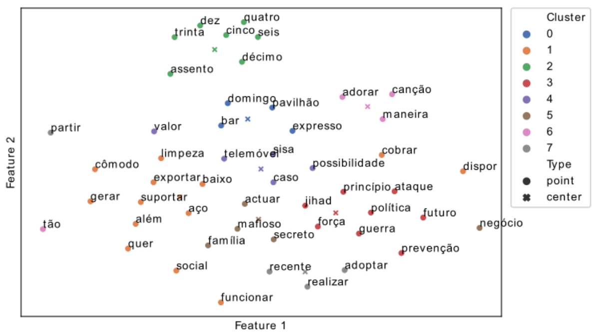

Figure 2: *Word clusters developed for patients diagnosed with psychosis on Task 6 using Latent Content Analysis and KMeans clustering. Manifold TSNE was used to visualize high-dimensional data in two dimensions.*

2.3.3. Vector Unpacking's Features

This method also has as its foundation the work of Rezaii et al. [14]. Once again, we used word embeddings computed using *Word2Vec* [30] from *CETEMPúblico* [29] corpus and average out the embeddings of groups of words.

The various word embeddings and corresponding group embeddings are then fed through the neural network as input and expected output, respectively. The neural network is composed of two layers, the input, and output layer, connected as displayed in Figure 3. The neural network must minimize the sum of squared errors by updating the various word weights. At any time, if the weights fall beneath a certain threshold defined through the following Equation 1, the weight value is set as 0.

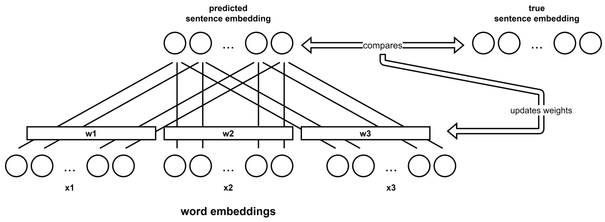

Figure 3: *Structure of Rezaii's Neural Network.*

$$\frac{iteration\,number}{\tau \times max\{iterations\}} \tag{1}$$

with τ being a constant with a value of 100, achieved through experimentation.

When the neural network has finished updating the weights and minimized the sum of squared errors, we extract as features for the classification: (1) the number of epochs for which the network had to run; (2) the ratio of weights set as 0 from the total number of weights trained; (3) the lowest loss score achieved by the network throughout its execution; and (4) the highest cosine difference achieved between the expected and achieved outputs throughout its execution.

This technique allows for the measurement of the semantic density of the discourse. Literature suggests that patients diagnosed with psychosis have a noticeably low semantic density, consequently, it was expected that more of the weights of the neural network would be set as zero for these patients.

2.3.4. Word Graph's Features

This technique allows for the assessment of topological structures of discourse [18, 9, 21, 13]. A connected and directed graph is created from transcribed speech samples where each node represents a given word, and each link represents temporal connections between words/nodes. From the word graph, topological structures can be evaluated and compared to others to identify possible deviations.

From word graphs we extracted: (1) number of nodes; (2) number of edges; (3) number of nodes in largest connected component; (4) number of nodes in largest strong connected component; (5) probability of the largest strongly connected component occurring; (6) number of repeated edges; (7) number of parallel edges; (8) size and number of cycles; (9) average total degree; and (10) the diameter. These features have been shown to relate well with speech attributes such as complexity, connectedness, and recurrence. Coincidentally, these speech attributes serve as good markers for psychosis identification in clinicians' diagnoses.

2.3.5. Sentiment Analysis' Features

The models and data available for sentiment analysis of European Portuguese are highly limited. Most of the data available pertain to social network posts, which differ greatly from the data achieved through our protocol, both in topic and in structure. For this reason, our team decided on a simpler technique. We used *SentiLex-PT* [31], a sentiment lexicon that maps lemmas and inflected forms of European Portuguese to -1 or 1, with

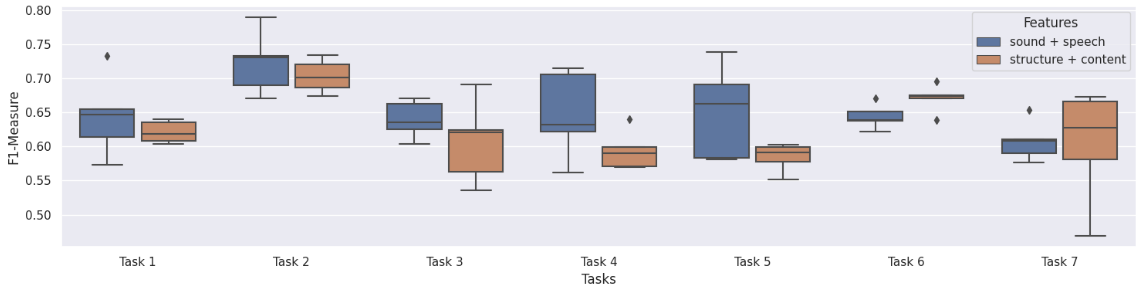

Figure 4: *Box plot of the results achieved with sound and speech features against the results achieved with structural and content features in each task.*

-1 being generally a negative token and +1 a positive token. This approach although simple has its limitations since it does not consider the sentence's structure and context.

Nonetheless, in an attempt to measure the valence of discourse, which has been shown to differ in patients diagnosed with psychosis due to their speech apathy [11], we used *SentiLex-PT*. We extracted both the number of tokens that were matched with the ones present on the lexicon and the average valence of the words matched.

3. Results

The final size of the First European Corpus for Psychosis Identification is limited, even after our efforts for its expansion. For this reason, when developing our models, we employed leave-one-out cross-validation, as to prevent model overfitting. Our team explored various well-known techniques for developing classification models such as Support Vector Machines, K-Nearest Neighbours, Decision Trees, Random Forests, and Feed-forward Multi-Layer Neural Networks. Lastly, we also explored variations in the values of the hyperparameters for the mentioned techniques.

To evaluate the models developed, we used the F1-Measure, defined as the harmonic mean of the precision and recall. The F1-Measure is considered a robust measurement of model quality, since it takes into consideration two different metrics, behaving well in both balanced and imbalanced datasets.

With the speech fluency, time, and acoustic features mentioned in section 2.2 the best score achieved was *0.7901*. This score was acquired by a Random Forest composed of 75 trees, with the entropy criterion, and a max depth of 32 for each one of the trees developed. The model was developed from the corpus recordings and transcriptions for Task 2.

Regarding the coherence, structural and content features mentioned in section 2.3 the best score achieved was *0.7340*. This score was achieved by a Multi-Layer Neural Network with two hidden layers, with 100 and 50 neurons respectively, using the hyperbolic tangent as their activation function, a constant learning rate of 0.005, and a maximum of 3000 iterations. The model was developed from the corpus transcriptions for Task 2.

When comparing these feature sets on a task level, as seen in Figure 4, it is noticeable that the structural and content features achieved undeniably better results than sound and speech features on Tasks 6 and 7. Nevertheless, as seen from Figure 5, sound and speech features achieved slightly better results than the structural and content features.

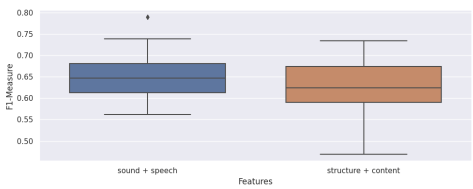

Figure 5: *Box plot of the results achieved with sound and speech features and with structural and content features.*

4. Discussion and Conclusions

The results obtained through fluency, time, and acoustic features were generally better than the results obtained through coherence, structural, and content features as revealed in section 3. This suggests that, at least for the current protocol, the first hold more power over the subjects' classification than the second.

However, the difference is small, and these types of features should not be discarded and should still be explored to access their full potential. A score of *73.40%* is far from what is expected from highly impactful domains such as mental disorder diagnosis. Nonetheless, continuous improvements in these domains could be helpful in early prevention efforts by flagging potential converters. The score achieved shows that classification is feasible, being well above random classification.

In Tasks 6 and 7, the structural and content features achieved slightly better results than sound and speech features. This may be since both of these tasks rely more on the subject and its interpretation of somewhat more negative images. Literature regarding patients with psychosis refers that there is some tendency from these patients to be more apathetic, and have less awareness regarding negative perceptions, which might provide some clarification on our results.

Due to space limitations, it was not possible to reveal further results and their analyses, such as the impact of the models developed on non-extended versions of the corpus.

In future work, our team will explore other methods for the extraction of more content features such as the use of web-scraped valence data to fine-tune a RoBERTa transformer to improve the current sentiment analysis described in subsection 2.3.5. Moreover, future work should also explore more methods for the extraction of content features from transcriptions, such as the extraction of the Level of Committed Belief, and possible alterations to the corpus acquisition protocol that incentivize subjects to speak more and more freely during their tasks.

5. Acknowledgments

This work was supported by national funds through *Fundação para a Ciência e a Tecnologia* (FCT) under project UIDB/50021/2020. We also thank the staff from Unidade de Saúde Mental de Oeiras - Centro Hospitalar de Lisboa Ocidental, for their help and support during the patients' corpus recording process.

6. References

[1] S. L. James, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017," *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, 2018.

[2] R. Coentre and P. Levy, "Early intervention in psychosis in portugal: Where are we?" *Acta medica portuguesa*, vol. 33, pp. 540–542, 9 2020.

[3] World Health Organization. (2021) Mental Disorders. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mental-disorders

[4] T. R. Insel, "Rethinking schizophrenia," *Nature*, vol. 468, pp. 187–193, 2010. [Online]. Available: https://doi.org/10.1038/nature09552

[5] E. J. Clemmer, "Psycholinguistic aspects of pauses and temporal patterns in schizophrenic speech," *Journal of Psycholinguistic Research*, vol. 9, pp. 161–185, 1980. [Online]. Available: https://doi.org/10.1007/BF01067469

[6] M. Forjó, H. S. Pinto, and A. Abad, "Contributions towards the possible identification of psychosis through speech processing in portuguese," Master's thesis, Instituto Superior Técnico - Universidade de Lisboa, 2021.

[7] B. Elvevåg, P. W. Foltz, M. Rosenstein, and L. E. Delisi, "An automated method to analyze language use in patients with schizophrenia and their first-degree relatives," *Journal of neurolinguistics*, vol. 23, pp. 270–284, 5 2010. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/20383310https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2850213/

[8] D. Iter, J. Yoon, and D. Jurafsky, "Automatic detection of incoherent speech for diagnosing schizophrenia," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 136–146.

[9] N. B. Mota, N. A. P. Vasconcelos, N. Lemos, A. C. Pieretti, O. Kinouchi, G. A. Cecchi, M. Copelli, and S. Ribeiro, "Speech graphs provide a quantitative measure of thought disorder in psychosis," *PloS one*, vol. 7, pp. e34 928–e34 928, 2012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/22506057https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322168/

[10] K. McManus, E. K. Mallory, R. L. Goldfeder, W. A. Haynes, and J. D. Tatum, "Mining twitter data to improve detection of schizophrenia," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2015, pp. 122–126, 3 2015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/26306253https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4525233/

[11] E. S. Kayi, M. Diab, L. Pauselli, M. Compton, and G. Coppersmith, "Predictive linguistic features of schizophrenia," *arXiv preprint arXiv:1810.09377*, 2018.

[12] G. Bedi, F. Carrillo, G. A. Cecchi, D. F. Slezak, M. Sigman, N. B. Mota, S. Ribeiro, D. C. Javitt, M. Copelli, and C. M. Corcoran, "Automated analysis of free speech predicts psychosis onset in high-risk youths," *npj Schizophrenia*, vol. 1, no. 1, pp. 1–7, 2015.

[13] N. B. Mota, M. Copelli, and S. Ribeiro, "Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance," *npj Schizophrenia*, vol. 3, p. 18, 2017. [Online]. Available: https://doi.org/10.1038/s41537-017-0019-3

[14] N. Rezaii, E. Walker, and P. Wolff, "A machine learning approach to predicting psychosis using semantic density and latent content analysis," *npj Schizophrenia*, vol. 5, p. 9, 2019. [Online]. Available: https://doi.org/10.1038/s41537-019-0077-9

[15] G. R. Kuperberg, P. K. McGuire, and A. S. David, "Reduced sensitivity to linguistic context in schizophrenic thought disorder: evidence from on-line monitoring for words in linguistically anomalous sentences." *Journal of abnormal psychology*, vol. 107, p. 423, 1998.

[16] C. M. Corcoran, F. Carrillo, D. Fernández-Slezak, G. Bedi, C. Klim, D. C. Javitt, C. E. Bearden, and G. A. Cecchi, "Prediction of psychosis across protocols and risk cohorts using automated language analysis," *World psychiatry : official journal of the World Psychiatric Association (WPA)*, vol. 17, pp. 67–75, 2 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29352548https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5775133/

[17] S. Just, E. Haegert, N. Kořánová, A.-L. Bröcker, I. Nenchev, J. Funcke, C. Montag, and M. Stede, "Coherence models in schizophrenia," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 2019, pp. 126–136.

[18] T. J. Spencer, B. Thompson, D. Oliver, K. Diederen, A. Demjaha, S. Weinstein, S. E. Morgan, F. Day, L. Valmaggia, G. Rutigliano, A. D. Micheli, N. B. Mota, P. Fusar-Poli, and P. McGuire, "Lower speech connectedness linked to incidence of psychosis in people at clinical high risk," *Schizophrenia Research*, vol. 228, pp. 493–501, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920996420304588

[19] M. Alpert, A. Kotsaftis, and E. R. Pouget, "Speech fluency and schizophrenic negative signs," *Schizophrenia Bulletin*, vol. 23, pp. 171–177, 1997.

[20] G. Gosztolya, A. Bagi, S. Szalóki, I. Szendi, and I. Hoffmann, "Making a Distinction Between Schizophrenia and Bipolar Disorder Based on Temporal Parameters in Spontaneous Speech," in *Proc. Interspeech 2020*, 2020, pp. 4566–4570. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-0049

[21] N. B. Mota, R. Furtado, P. P. C. Maia, M. Copelli, and S. Ribeiro, "Graph analysis of dream reports is especially informative about psychosis," *Scientific Reports*, vol. 4, p. 3691, 2014. [Online]. Available: https://doi.org/10.1038/srep03691

[22] A. Pompili, R. Solera-Ureña, A. Abad, R. Cardoso, I. Guimaraes, M. Fabbri, I. P. Martins, and J. Ferreira, "Assessment of parkinson's disease medication state through automatic speech analysis," *arXiv preprint arXiv:2005.14647*, 2020.

[23] C. Carvalho and A. Abad, "TRIBUS: An end-to-end automatic speech recognition system for European Portuguese," in *Proc. IberSPEECH 2021*, 2021, pp. 185–189.

[24] G. A. Miller and J. A. Selfridge, "Verbal context and the recall of meaningful material," *The American journal of psychology*, vol. 63, no. 2, pp. 176–185, 1950.

[25] C. E. Shannon, "Prediction and entropy of printed english," *Bell system technical journal*, vol. 30, no. 1, pp. 50–64, 1951.

[26] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. [Online]. Available: https://nlp.stanford.edu/pubs/qi2020stanza.pdf

[27] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[28] audEERING GmbH. (2021) openSMILE Python. [Online]. Available: https://audeering.github.io/opensmile-python/s

[29] P. A. Rocha and D. Santos, "Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa," *quot; In Maria das Graças Volpe Nunes (ed) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)(Atibaia SP 19-22 de Novembro de 2000) São Paulo: ICMC/USP*, 2000.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[31] P. Carvalho and M. J. Silva, "Sentilex-pt: Principais características e potencialidades," *Oslo Studies in Language*, vol. 7, no. 1, 2015.