IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 4, 2019, accepted March 19, 2019, date of publication April 1, 2019, date of current version April 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908620

A Machine Learning Approach for the Automatic Classification of Schizophrenic Discourse

HÉCTOR ALLENDE-CID[1], JUAN ZAMORA[2], PEDRO ALFARO-FACCIO[3], AND MARÍA FRANCISCA ALONSO-SÁNCHEZ[4]

[1]Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Valparaíso 2340025, Chile
[2]Instituto de Estadística, Pontificia Universidad Católica de Valparaíso, Valparaíso 2340025, Chile
[3]Instituto de Literatura y Ciencias del Lenguaje, Pontificia Universidad Católica de Valparaíso, Viña del Mar 2340025, Chile
[4]Centro de Investigación del Desarrollo en Cognición y Lenguaje, Universidad de Valparaíso, Viña del Mar 2391415, Chile

Corresponding author: Juan Zamora (juan.zamora@pucv.cl)

The work of H. Allende-Cid was supported by the CONICYT-FONDECYT, Government of Chile, under Grant 11150248. The work of J. Zamora was supported by the CONICYT-FONDECYT, Government of Chile, through the Postdoctoral Project under Grant 3180689.

ABSTRACT Schizophrenia is a chronic neurobiological disorder whose early detection has attracted significant attention from the clinical, psychiatric, and also artificial intelligence communities. This latter approach has been mainly focused on the analysis of neuroimaging and genetic data. A less explored strategy consists in exploiting the power of natural language processing (NLP) algorithms applied over narrative texts produced by schizophrenic subjects. In this paper, a novel dataset collected from a proper field study is presented. Also, grammatical traits discovered in narrative documents are used to build computational representations of texts, allowing an automatic classification of discourses generated by schizophrenic and non-schizophrenic subjects. The attained results showed that the use of the proposed computational representations along with machine learning techniques enables a novel and precise strategy to automatically detect texts produced by schizophrenic subjects.

INDEX TERMS Applied machine learning, natural language processing, schizophrenia.

I. INTRODUCTION

Schizophrenia is a chronic neurobiological disorder with recurrent tendency and wide heterogeneity of positive, negative and mood symptoms. Some of the symptoms associated with the disease are: delusions, hallucinations, catatonic or disorganized behavior, apathy, reduced thought fluidity, disperse and unproductive language, and difficulty with goal oriented behaviors [1]. Besides the symptoms described before, Schizophrenia also involves alterations in executive function, psycho-motor speed and social skills [2]. Another relevant feature is the impediment that these symptoms generate in social, occupational and daily life activities [2]. Although the relatively low rates of schizophrenia incidence worldwide - around 15 per 100.000 per year- it is considered a devastating pathology due to the impact in the community participation functionality [3].

Communication in Schizophrenia is heterogeneous and the descriptions are usually subjective and unspecific [4]. In fact, most of the Schizophrenic language research assesses general aspects of comprehension and production of oral and written speech. This seems paradoxically based on the fact of the discursive, pragmatic, syntactic, morphological and phonological particularities of language are useful to find differences among groups according to their neurolinguistics conditions. Therefore, although schizophrenia is not language pathology, the speech of these patients could serve as a distinctive marker. In fact, particularly in schizophrenia, several research have been made in order to decrease the false positives in the diagnostic process [5]. In this context, the use of computational methods could allow to implement this analysis automatically and objectively. Hence, computational linguistics, specifically latent semantic analysis, has demonstrated to successfully index the thought disorders according to lexical co-occurrence in texts [6].

Computer intelligence based on statistical learning theory has brought a powerful tool for automatically grasping hidden patterns in data. This is, intelligent algorithms use features thoroughly extracted from the phenomenon to learn a specific task, for example the categorization of a subject in the control

The associate editor coordinating the review of this manuscript and approving it for publication was Damon Lamar Woodard.

of the patient group, and then they perform this task over a new datum. Specifically for text analysis, classification methods such as Support Vector Machines enable an accurate discrimination of texts represented as bags of features [7]. Additionally, some other methods such as Bayes Nets not only allow an adequate classification but also provide a theoretical framework quite useful to unravel the traits that determine most the classification boundary as described in [8]. Some works in which these models (along with some others) have been successfully employed in the diagnosis of mental diseases are [9], [10].

This work poses that narrative texts produced by subjects under the observation of a carefully designed task can be analyzed using Machine Learning in order to identify the presence of Schizophrenic traits. To accomplish this aim, novel computational feature representations for narrative texts based on Part–of–Speech tagging (POS) are proposed and then two automatic methods are employed to assess its discriminative power.

A. ORGANIZATION OF THIS DOCUMENT

This work is structured as follows: In the next section we discuss previous approaches in which the automatic classification of schizophrenia has been addressed. In Section III we present the data collection procedure. Following in section IV, a probabilistic framework is used in order to assess the discriminative power of the linguistic features that will be employed for the classification task. Subsequently, two low dimensional document representations, namely the POS and the Meta-POS, are introduced in section V. Also in that section, preliminary experiments with an interpretable model are performed and a discussion about how the classification task could be improved is conducted. Then, a more thorough experimentation that considers state-of-the-art techniques together with the proposed representations is performed and the attained results are discussed in detail. The last section is devoted to present the final conclusions and future work.

II. STATE OF THE ART

Schizophrenia is a mental disease that has intrigued psychiatrist and scientist in general for a long time. Due to the severity and impact of this illness it has drawn the attention of scientists from different areas of research, such as biomedical engineering and computational linguistics, since an intervention in an early stage seems to favorably influence the short-term illness course. One source of information related with this work consists in the manifestation of cognitive impairments as a deficit in the verbal-working-memory as mentioned in [11].

A modern approach to exploit this source of evidence comes from the Computational Intelligence community, that with automated algorithms has tried to find significant patterns that differentiate people that suffers from Schizophrenia. Among the techniques used for this Pattern Recognition process is Machine Learning. This area explores the study and construction of algorithms that can learn from and make predictions on data. These algorithms overcome following strictly static program instructions by making data driven predictions or decisions, through building a model from sample inputs. One can see Machine Learning algorithms as Pattern Recognition systems that discover the underlying patterns that allow to classify or infer on new unseen data, thus making them inference models.

The study of Schizophrenia from the Machine Learning and Statistical Analysis community has been very active in the last decades. One of the most studied approaches has been centered in Biomedical signals, such as Electro Encelography (EEG) signals and Magnetic Resonance Images (MRI). In [12] Multivariate machine learning methods are used to classify groups of schizophrenia patients and controls using structural magnetic resonance imaging (MRI). The authors hypothesized that brain measures would classify groups, and that increased likelihood of being classified as a patient using regional brain measures would be positively related to illness severity, developmental delays and genetic risk. The authors state that Schizophrenia and control groups can be well classified using Random Forest and anatomic brain measures (achieving 73.7% accuracy), and brain-based probability of illness has a positive relationship with illness severity and a negative relationship with developmental delays/problems and CNV-based risk. In [13] the authors evaluate the overall reliability of neuroimaging-based biomarkers, conducting a comprehensive literature search to identify all studies that used multivariate pattern recognition to identify patterns of brain alterations that differentiate patients with schizophrenia from healthy controls. A bivariate random-effects meta-analytic model was implemented to investigate the sensitivity and specificity across studies as well as to assess the robustness to potentially confounding variables. More recently, in [14] a 2-stage Stacked AutoEncoder based architecture is proposed for classification of normal versus Schizophrenic subjects from functional MRI data. First, an auto encoder network is employed to generate vector representations of each brain region (previously filtered from the identified active voxels). Then, these vectors along with the participant labels (Schizophrenic and non-Schizophrenic) are passed to a Support Vector Machine as train data for the binary classification task. The authors attain an accuracy of over 90% with the proposed Deep Learning framework.

On the other hand, there are several studies that deal with EEG signals. In [15] electroencephalogram (EEG) signals of 13 schizophrenic patients and 18 age-matched control participants were analyzed with the objective of classifying the two groups. For each case, multi-channels (22 electrodes) scalp EEG were recorded. Several features including autoregressive (AR) model parameters, band power and fractal dimension were extracted from the recorded signals. Leave-one (participant)-out cross validation was used to have an accurate estimation for the separability of the two groups. Boosted version of Direct Linear Discriminant Analysis (BDLDA) was selected as an efficient classifier which applied on the extracted features, obtaining 87.51% in accuracy. In [16] the authors propose a two stage procedure for analysis and classification of electroencephalogram (EEG) signals for twenty schizophrenic patients and twenty age-matched control participants. For each case, 20 channels of EEG were recorded. First, the more informative channels were selected using the mutual information techniques. Then, genetic programming was employed to select the best features from the selected channels. Several features including autoregressive model parameters, band power and fractal dimension were used for the purpose of classification. Both linear discriminant analysis (LDA) and adaptive boosting (Adaboost) were trained using 10-fold cross validation to classify the reduced feature set and a classification accuracy of 85.90% and 91.94% was obtained by LDA and Adaboost, respectively. There are several other works that propose Machine Learning based methods for feature selection and reduction [17], [18] for the same classification task.

Studies that aim to make relations between language and Schizophrenia using Machine Learning models are scarce. In [19] a work that reports the first results of a simulation of language pathology in schizophrenia is presented. Using DISCERN, a subsymbolic model of story understanding and recall, the impact of different simulated lesions hypothesized to underlie schizophrenia is investigated. In response to excessive connection pruning, the model reproduces symptoms of delusions and disorganized language seen in schizophrenia, as well as the reduced output seen in compensated later states of the disorder. In [20] the authors propose a work that aims to capture the link between biology and schizophrenic symptoms using also DISCERN. Competing illness mechanisms proposed to underlie schizophrenia are simulated in DISCERN, and are evaluated at the level of narrative language, i.e. the same level used to diagnose patients. The result is the first simulation of abnormal storytelling in schizophrenia, both in acute psychotic and compensated stages of the disorder. The authors of [21] explore potential linguistic markers of Schizophrenia using the tweets of self-identified schizophrenia sufferers, and describe several natural language processing (NLP) methods to analyze the language. The authors examine how these signals compare with the widely used LIWC categories for understanding mental health and provide preliminary evidence of additional linguistic signals that may aid in identifying and getting help to people suffering from schizophrenia. In [22] the authors state that prominent formal thought disorder, expressed as unusual language in speech and writing, is often a central feature of Schizophrenia. Thirty-six patients with DSM-IV criteria chronic Schizophrenia provided a page of writing (300-500 words) on a designated topic. Writing was examined by automated text categorization and compared with non-psychiatrically ill individuals, investigating any differences with regards to lexical and syntactical features. Computerized methods used included extracting relevant text features, and using Machine Learning techniques to induce mathematical models distinguishing between texts belonging to different categories. Observations indicated that automated methods distinguish schizophrenia writing with 83.3% accuracy. Results reflect underlying impaired processes including semantic deficit, independently establishing connection between primary pathology and language. In [23], through the examination of the performance on an on-line word-monitoring task, the use of linguistic context in positively thought-disordered (TD) schizophrenics was investigated. In [24], the authors tested the hypothesis that schizophrenia patients show impairments in building up context within sentences because of abnormalities in combining semantic with syntactic information. Recently, in [25] an attempt to detect individuals with schizophrenia from their profiles and posting history in Twitter is made. A total of 28 features are extracted and used to train several automatic classifiers. Some of the Finally, the best result attained over 20% of the data in terms of F1 measure is 0.8. There are several other studies that study the effect of Schizophrenia in language [26], [27].

III. DATA DESCRIPTION AND COLLECTION PROCEDURE

A. CORPUS

The corpus recolected was composed by one hundred eighty nine texts ($n = 189$) compiled through three oral narrative tasks.

Thirty nine texts were acquired from thirteen patients with the diagnosis of chronic undifferentiated schizophrenia, according to the DSM IV and recruited from a rehabilitation center. The inclusion criteria were to be behaviorally compensated and with stable medication. The ages range were between 19 and 74 years old in order to represent the heterogeneity of the population and their speech abilities through life span.

Furthermore, considering a significance of .05, a statistic power of .95 and a size effect of 1.67, the sample size required was 9 subjects. Moreover, all the subject belonged to the lower socioeconomic status in accordance to the epidemiological description.

The rest of the corpus was composed by one hundred and fifty stories, produced by fifty healthy volunteers without history of language or mental illness. Their ages ranges were between 20 and 30 years old and, as the law require they had at least 12 years of education. In spite of the obvious difference between the age ranges of both groups, the group of healthy subjects represents the average speaking person from a normal population and thus we consider this a valid contrast in order to assess the discrimination power of the proposed method.

B. PROCEDURE

Three stories were visually presented on a sequence to the participants. Each story had the same structure, and was divided on presentation of the character with the context and the personal motivation, an initial event with the triggering and the consequences, a development of the story with an action plan, a final event with an initial suggestion, a counterattack and a climax, and finally a resolution of the story with the new context and state. This decision is due to the fact that we sought to have homogeneous data through the same linguistic task.

The stories were illustrated in a book format. Each frame present an item of the story structure with a big and colored drawing. Every participant were asked to tell the three stories in a quiet room where all the narratives were recorded on digital high definition media. There was no time limit and the participants were able to review several times the sequence to reduce the memory effect. The audio were manually transcribed using a orthographic transcription. The computational analysis was made on these digitalized texts. According to the Helsinki statement, all the participants signed an informed consent approved by the ethics committee.

IV. ANALYSIS OF LINGUISTIC MARKERS FOR THE REPRESENTATION OF TEXT DOCUMENTS

The collected data consists of 3 datasets made up by the corresponding narrations of each story types of participants (see Table 1). As the participants in each experiment were the same, each dataset contains the same number of texts produced by the same participants.

TABLE 1. Number of participants in each group for each story document set.

	number of participants	
Story	**Control**	**Experimental**
A	50	13
B	50	13
C	50	13

Each oral narration was transcribed, digitalized and processed by extracting and counting Part-Of-Speech tags. The total number of features extracted by following the previous step is 163. All these tags denote the kind of linguistic information that the automatic classifiers will make use of. Additionally and in order to enhance the linguistic and clinical interpretation, we attempt to exploit the information contained within this set of features by quantifying the extent of dependence between each feature and the participant group, i.e. text produced by schizophrenic and non-schizophrenic individuals. All the procedures described in this section were performed over the collected story narrations created by the participants under study.

A. LINGUISTIC FEATURE GENERATION

Consider two sets of POS-tags differenced each other only by the narrowness of the contained linguistic characteristics. The first one, denoted by Ω contains specific tags, for instance *Noun Common Masculine Singular*. The latter, denoted by Ψ contains open POS tags such as *Noun*, *Verb* and *Determiner* among others. We will denote the elements of this latter set, as meta-POS-tags. It is also possible to induce a total order over Ψ by applying a lexicographic order over the tag names, then each of the elements of this set can be mapped onto the interval $\{1, 2, \ldots, |\Psi|\} \subset \mathbb{Z}$ by an ad-hoc function $\lambda : \Psi \rightarrow \{1, 2, \ldots, |\Psi|\}$ that sorts the tags by name and assigns and index to each one. Let $\phi : \Omega \rightarrow \{1, 2, \ldots, |\Psi|\}$ be a subjective function that maps each narrow tag to the index associated to a meta-POS-tag category in Ψ. For instance,

$$\phi(\text{Noun Common Masculine Singular}) = \lambda(\text{Noun})$$

or

$$\phi(\text{Determiner Indefinite Femenine Singular}) = \lambda(\text{Determiner}).$$

In this work, we attempt to reduce a document representation to a bag-of-POS-tags, i.e. a tuple of measurements of the presence of items in Ω within a digitalized written text. Under this concept, it is possible to represent initially each document d as a set of tuples

$$x' = \{\langle t, f_d(t)\rangle | t \in \Omega\}$$

where $f_d(t)$ denotes the (raw or normalized) number of terms within document d POS-tagged as $t \in \Omega$. Finally, by exploiting the total order induced onto Ψ and using the mapping ϕ defined above, this set can be converted to a vector $\mathbf{x} \in \mathbb{Z}^{|\Psi|}$, whose value associated to POS tag i is given by

$$\mathbf{x}_i = f_i(t) \tag{1}$$

Alternatively, the value of this document vector associated to the i-th meta POS tag is given by

$$\mathbf{x}_i = \sum_{\substack{\langle t, f_d(t)\rangle \in x' \\ \text{s.t. } \phi(t)=i,\ i\in\{1,2,\ldots,|\Psi|\}}} f_d(t) \tag{2}$$

B. ANALYSIS OF CAUSAL RELATIONSHIP BETWEEN EACH POS-TAG AND THE PARTICIPANT GROUP

The dependence of the type of text in terms of each POS feature was explored by modeling their relationships by using a probabilistic framework. In order to do this, each document is treated as a bag of POS-tags and it is also assumed that the position of each tag occurs independently from the others. Therefore, all POS-tags extracted from the available document sets are considered as features that represent every document. Since the written texts collected in this experimentation have a relatively homogeneous length, only raw POS tag frequencies are considered.

As the category to which each of the text belongs is known, namely texts written by schizophrenic and non-schizophrenic participants, it is possible to estimate the probability of observing any feature $t \in \Psi$ in a document vector $\mathbf{x}$ given that its writer was a schizophrenic or control individual. This is $P(x_1, x_2, \ldots, x_{|\Psi|}|C_{\text{neg}})$ or $P(x_1, x_2, \ldots, x_{|\Psi|}|C_{\text{pos}})$ respectively.

$$\begin{aligned} P(\mathbf{x}|C_{\text{pos}}) &= P(x_1, x_2, \ldots, x_{|\Psi|}|C_{\text{pos}}) = \Pi_{t\in\Psi} P(t|C_{\text{pos}}) \\ P(\mathbf{x}|C_{\text{neg}}) &= P(x_1, x_2, \ldots, x_{|\Psi|}|C_{\text{neg}}) = \Pi_{t\in\Psi} P(t|C_{\text{neg}}) \end{aligned} \tag{3}$$

TABLE 2. Sub-collections built from the three original stories.

training mix		testing	new dataset identifier	Dim. Meta-POS	Dim. POS	Dim. TF-IDF
A	B	C	1	12	163	2893
A	C	B	2	12	159	2656
B	C	A	3	12	160	2816

On the one hand, the probabilities $P(C_{\text{neg}})$ and $P(C_{\text{pos}})$ can be estimated from the data as the fraction of documents in each category. On the other hand, there are several options to estimate the probabilities $P(\mathbf{x}|C_{\text{neg}})$ and $P(\mathbf{x}|C_{\text{pos}})$, e.g. Maximum Likelihood or MAP estimates. The one chosen for this work consists in assuming a Multinomial distribution for the probability density functions above with parameters $P(t|C_{\text{neg}})$ and $P(t|C_{\text{pos}})$ for each $t \in \Psi$ respectively. That is

$$P(\mathbf{x}|C_{\text{pos}}) = \frac{\left(\sum_{i=1}^{|\Psi|} x_i\right)!}{x_1!x_2!\ldots x_{|\Psi|}!} \times \Pi_{t\in\Psi} P(t|C_{\text{pos}})^{x_{\lambda(t)}}$$

$$P(\mathbf{x}|C_{\text{neg}}) = \frac{\left(\sum_{i=1}^{|\Psi|} x_i\right)!}{x_1!x_2!\ldots x_{|\Psi|}!} \times \Pi_{t\in\Psi} P(t|C_{\text{neg}})^{x_{\lambda(t)}} \qquad (4)$$

The probabilities $P(t|C_{\text{pos}})$ and $P(t|C_{\text{neg}})$ are estimated from the tag frequencies within documents in each document category as follows:

$$P(t|C_{\text{pos}}) = \frac{\sum_{\mathbf{x}\in\mathcal{D}_{\text{pos}}} x_{\lambda(t)}}{\sum_{t'\in\Psi}\sum_{\mathbf{x}\in\mathcal{D}_{\text{pos}}} x_{\lambda(t')}}$$

$$P(t|C_{\text{neg}}) = \frac{\sum_{\mathbf{x}\in\mathcal{D}_{\text{neg}}} x_{\lambda(t)}}{\sum_{t'\in\Psi}\sum_{\mathbf{x}\in\mathcal{D}_{\text{neg}}} x_{\lambda(t')}} \qquad (5)$$

Finally, the posterior probability of a new document $\mathbf{z}$ for both categories will be given by

$$\begin{aligned}
P(C_{\text{pos}}|\mathbf{z}) &= P(\mathbf{z}|C_{\text{pos}}) \times P(C_{\text{pos}}) \\
&= \frac{\left(\sum_{i=1}^{|\Psi|} z_i\right)!}{z_1!z_2!\ldots z_{|\Psi|}!} \times \Pi_{t\in\Psi} P(t|C_{\text{pos}})^{z_{\lambda(t)}} \times P(C_{\text{pos}}) \\
P(C_{\text{neg}}|\mathbf{z}) &= P(\mathbf{z}|C_{\text{neg}}) \times P(C_{\text{neg}}) \\
&= \frac{\left(\sum_{i=1}^{|\Psi|} z_i\right)!}{z_1!z_2!\ldots z_{|\Psi|}!} \times \Pi_{t\in\Psi} P(t|C_{\text{neg}})^{z_{\lambda(t)}} \times P(C_{\text{neg}})
\end{aligned} \qquad (6)$$

C. MEASUREMENTS OF THE DISCRIMINATIVE POWER OF POS FEATURES

As an initial remark about the available data, as it is shown in Table 1, the document classes are unbalanced. Hence, to enable a correct interpretation of the results a SMOTE [28] sampling strategy is performed to tackle this issue. After the data is balanced, each pair of document sets are mixed an treated as a training dataset, i.e. stories A with B, A with C and B with C, and in each case the remaining document set is treated as a testing dataset. In other words, we perform 3-fold cross-validation, where each story is a fold. The generated datasets along with the identifiers with which they are named in the remainder of this work are shown in Table 2. Additionally, the dimensionality of the feature space obtained for representations 1, 2 and TF-IDF [1] are shown in the last three columns.

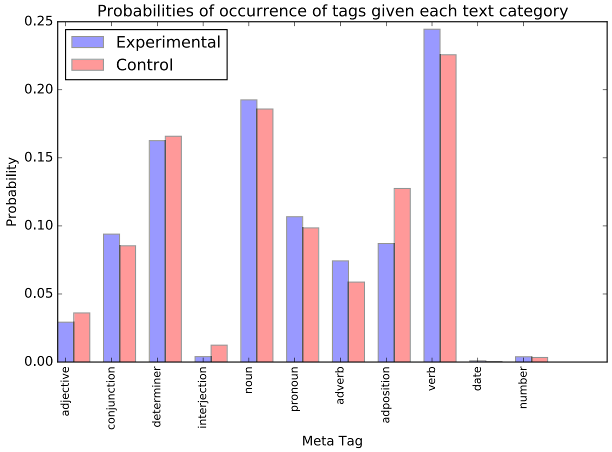

FIGURE 1. Probability of each meta POS tag conditioned to the type-of-written text class for dataset 1.

Additionally, we employ two sets of POS features to represent documents, originated from two different linguistic levels. First, highly specific linguistic POS features, e.g. Common singular noun (which characterizes words such as "activity" or "house") and secondly, less specific linguistic traits (Meta-POS features), e.g. Noun (which features words such as "helicopters" and "Saturday"). Hence, the analysis is going also to be conducted in a twofold way.

In order to assess the utility of the POS-tags (in both levels) extracted from each sub-collection, the model described in IV-B is fitted with the 3 different document collections described previously. The aim of using this strategy is twofold: First, to empirically quantify the discriminative power of the features in terms of both groups of participants. Second, gaining comprehension about which features have a greater impact in the discriminative process.

After fitting the model for each one of the three datasets in Table 2, posterior probabilities for each Meta-POS feature are computed by Eq. 5. The computation of these probabilities for the Meta-POS features can be performed from the POS-tag features (by applying the second law of probability) or straight from the Meta-POS-tag feature counts since both schemes lead to the same estimates. The values obtained are depicted as bar plots in Figures 1, 2 and 3.

[1] Term frequency - Inverse Document frequency vector representation

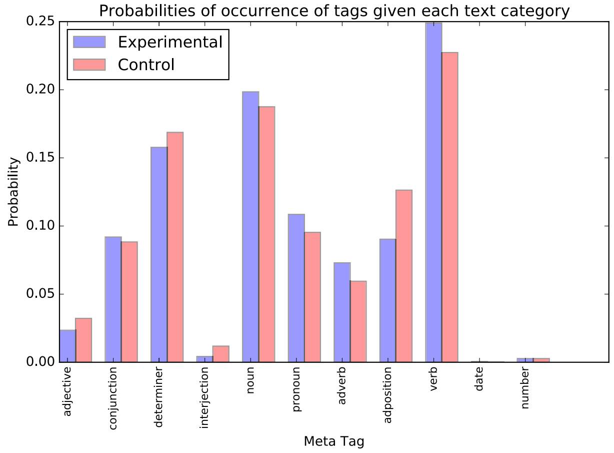

FIGURE 2. Probability of each Meta-POS tag conditioned to the type-of-written text class for dataset 2.

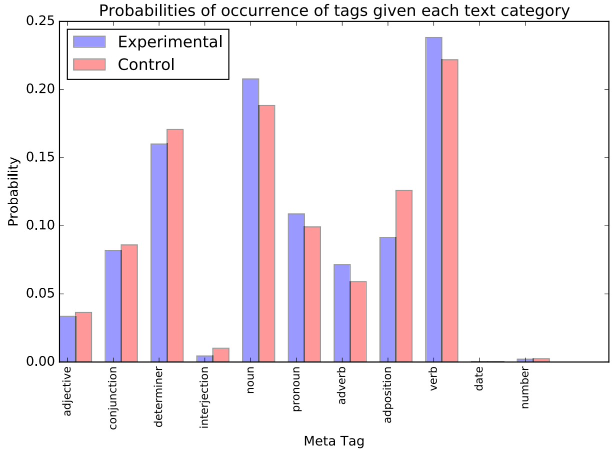

FIGURE 3. Probability of each meta POS tag conditioned to the type-of-written text class for dataset 3.

D. EVALUATION MEASURES

In order to evaluate the predictive performance of the classification algorithm with each feature representation over the three datasets, each pair of datasets is employed to train the model and the remaining one is used to test its performance. The class with maximum posterior probability in Eq. 6 is used to label each unseen document. In the training and testing steps Precision, Recall, the Harmonic Mean between them (F1) and the area under the ROC curve are computed. The Precision measure indicates the portion of texts identified as schizophrenic that effectively were generated by schizophrenic subjects. The Recall measure quantifies the portion of texts correctly identified as schizophrenic from the complete set of texts generated by schizophrenic subjects. The values of these two measures together with the F1 denote a better classifier as they get closer to 1.0. The area under the ROC curve measures the performance of a classification model based on the contrast between the True Positive and False Positive rates, and it denotes a better classifier as it gets closer to 1.0.

The values attained by the Naive Bayes model with each one of the two feature representations are shown in Tables 3 and 4.

E. DISCUSSION

The three figures show a very similar pattern for each feature. In spite of the fact that differences appeared between the two classes in all the features, those that exhibited higher contrasts are Verbs, Prepositions, Determiners and Pronouns. The discriminant function built with model 3 over the proposed linguistic features allows to separate between the two groups of narrative texts with scores above the 70% in F1 measure. Additionally, over all datasets the representations employed by the probabilistic classifier attained an Area-Under-the-ROC-Curve above (AUROC) 90%, which denotes a high probability that this classifier ranks a randomly chosen text written by a schizophrenic subject as positive than a randomly chosen text written by a control participant [29].

The evidence found in the previous analysis empirically demonstrates that, in the first place, almost all Meta-features

TABLE 3. Discriminative power of model 3 quantified in terms of average F1, Precision(P) and Recall(R) measures over Meta-POS-features using the normalized term frequency weighting scheme.

			training			*testing*		
Dataset	**AUROC**	**Class**	**F1**	**P**	**R**	**F1**	**P**	**R**
1	0.917(0.011)	Experimental	0.844(0.011)	0.800(0.014)	0.898(0.898)	0.748(0.010)	0.749(0.026)	0.750(0.750)
		Control	0.821(0.015)	0.885(0.018)	0.770(0.019)	0.744(0.020)	0.748(0.009)	0.743(0.039)
2	0.891(0.016)	Experimental	0.813(0.021)	0.798(0.023)	0.832(0.832)	0.782(0.008)	0.782(0.019)	0.783(0.783)
		Control	0.803(0.021)	0.826(0.025)	0.785(0.028)	0.780(0.014)	0.782(0.004)	0.779(0.025)
3	0.906(0.012)	Experimental	0.823(0.024)	0.831(0.031)	0.824(0.824)	0.852(0.023)	0.872(0.017)	0.838(0.838)
		Control	0.821(0.027)	0.829(0.035)	0.823(0.044)	0.860(0.015)	0.848(0.036)	0.876(0.024)

TABLE 4. Discriminative power of model 3 quantified in terms of average F1, Precision(P) and Recall(R) measures over POS-features using the normalized term frequency weighting scheme.

			training			*testing*		
Dataset	**AUROC**	**Class**	**F1**	**P**	**R**	**F1**	**P**	**R**
1	0.955(0.007)	Experimental	0.878(0.014)	0.838(0.017)	0.926(0.926)	0.779(0.005)	0.691(0.007)	0.892(0.892)
		Control	0.860(0.017)	0.919(0.017)	0.814(0.022)	0.703(0.010)	0.848(0.012)	0.600(0.015)
2	0.911(0.012)	Experimental	0.826(0.014)	0.800(0.016)	0.859(0.859)	0.890(0.013)	0.852(0.015)	0.932(0.932)
		Control	0.808(0.018)	0.849(0.020)	0.777(0.022)	0.879(0.013)	0.927(0.023)	0.838(0.019)
3	0.917(0.013)	Experimental	0.828(0.023)	0.787(0.018)	0.879(0.879)	0.862(0.009)	0.828(0.015)	0.901(0.901)
		Control	0.804(0.020)	0.866(0.036)	0.756(0.015)	0.848(0.012)	0.891(0.007)	0.811(0.019)

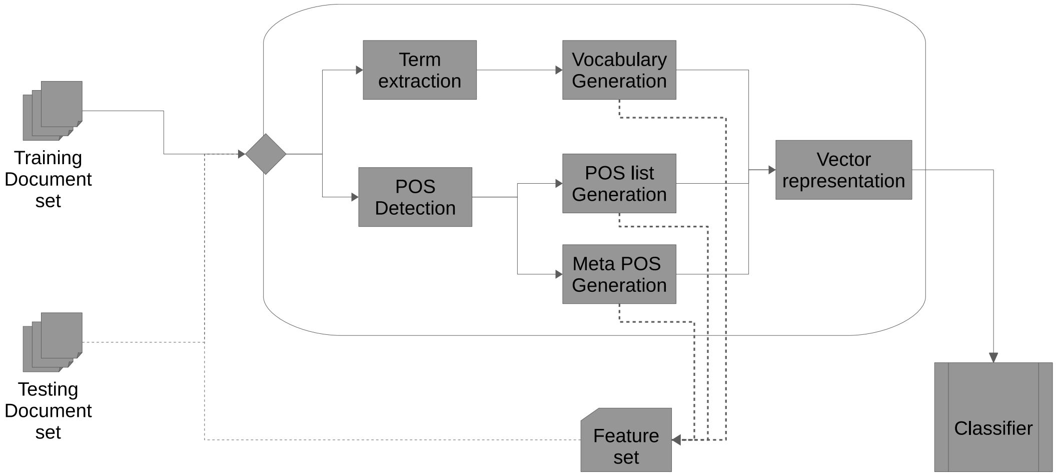

FIGURE 4. Processing stages applied for each document within a dataset.

showed a contrast in the values for their probability of appearance given the two document groups. In the second place, that a document representation built by using the POS-features (in both levels) allows to discriminate between both classes of narrative texts by using interpretable and simple classifiers, moving away the discrimination boundary from the random classification (as shown by the AUROC value) and attaining acceptable performance values over independent testing datasets.

V. IMPROVING THE CLASSIFICATION PERFORMANCE

The data employed consisted of three document collections generated as described in section III. In order to validate the proposed document characterization, three datasets are built by joining each pair of sub-collections as a training set and using the remaining portion for testing. In the following part of the section each dataset is going to be called as shown in table 2.

In contrast to the results shown in section IV-C, and in order to assess the precision within reach by using more computational power, four techniques are tested. The rationale behind this experimentation consists in exploiting at maximum the discriminative power of the selected features by using methods that build non-linear separation boundaries between the two classes in detriment of the interpretative power of the final solution. We suggest that an application that implements the expert system proposed in this work must sacrifice the interpretation by a reduction in the number of False Positive instances, i.e. texts written by schizophrenic subjects but identified as coming from a control individual.

A. COMPUTATIONAL REPRESENTATIONS

As depicted in Figure 4, each document is processed and represented by several features. Three different text representations are generated. The two novel linguistic representations presented in expressions 1 and 2, namely a POS feature and a Meta-POS-feature representation, and the standard TF-IDF vector representation commonly used for text classification [30].

B. METHODOLOGY

In order to assess the discriminative power of the algorithms, the training data was used together with each technique in a 3-Fold cross-validation procedure. Additionally, each cross-validation step was run 10 times in order to reduce the effect of the fold splits. At the end of each cross-validation step the performance attained over the testing fold and the evaluation set were registered. Then, the reported results for each dataset and each algorithm were computed by averaging the performances attained over the training and testing sets in each of the 3 Folds and also by averaging these results across the different runs. Additionally, the standard deviations are also reported.

C. PARAMETER TUNING

First and in order to better exploit each algorithm, a grid for their parameter values is defined. The combination of values in each cell of the grid is used to repeatedly train and test both algorithms only over each training dataset by following a 3-Fold cross-validation strategy. In every case, the combination that allowed to attain the best performance in terms of **F1** measure is selected. Once the parameter values are fixed and in pursuance of lessen any effect of an arbitrary split of the data during each training step, several runs are performed and within each one a random shuffle of the data is applied. Moreover in each run, a 3-Fold cross-validation is performed over the training data and at the end of each fold the testing dataset is presented to the classifier. Performance measures attained over training and testing data are averaged along the different folds and runs.

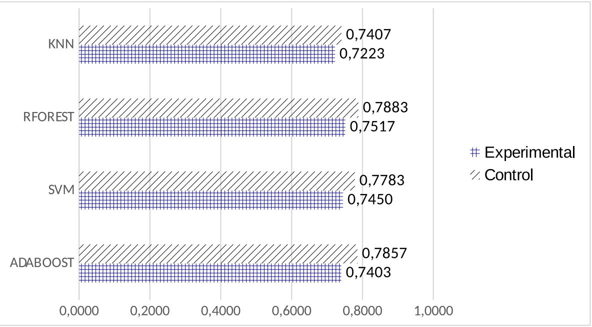

FIGURE 5. Performance attained by the less interpretable classifiers over the Meta feature document representation.

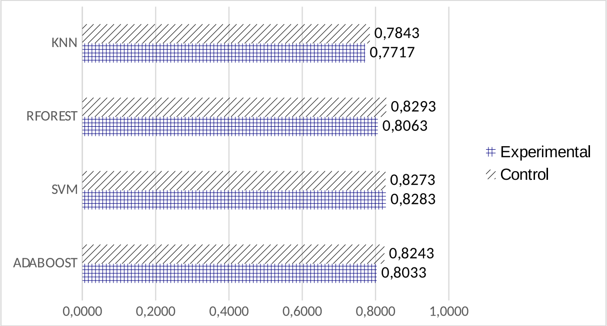

FIGURE 6. Performance attained by the less interpretable classifiers over the POS-tag feature document representation.

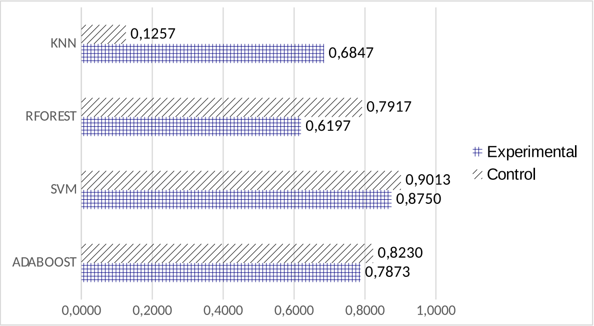

FIGURE 7. Performance attained by the less interpretable classifiers over the TF-IDF document representation.

D. DISCUSSION OF THE RESULTS

The attained results by the four classifiers over each class are shown in Figures 5, 6 and 7. Lighter color bars shown F1 values over the texts produced by Non-Schizophrenic subjects.

An interesting aspect to observe is that as the dimensionality of the feature space increases (see the dimensionality after generating each document representation in Table 2), the performance attained by methods designed for dealing with the curse of dimensionality, such as the SVM, also improves. The opposite behavior is shown for the KNN, which is a distance-based algorithm, where the performance decreases. The Meta and POS feature representations employ approximately a 0.4% and a 6% respectively of the space required by the TF-IDF representation. This fact also impacts on the execution time spent by each algorithm.

Even when the dimensionality of the feature space presents a high variation across the different representations, the lower ones, i.e. Meta and POS features representation, allows to attain comparable performance values in comparison to the TF-IDF . This suggests that, besides its simplicity, the more general document characterization proposed in this work enables a powerful discrimination between the distinguished classes. Additionally, the proposed linguistic characterizations not only allow a good discrimination but also their features are language-independent since only need Part-of-Speech tagging procedure.

VI. CONCLUSIONS

In this work, two novel document feature representations are proposed for the automatic identification of narrative texts produced by schizophrenic and control participants. Moreover, the study is conducted over real data gathered by specialists and never used before for this task. As mentioned in the state of the art, computational methods to identify schizophrenic have been proposed in the literature, but as far as we know, any of them addresses the problem from a linguistic and textual approach.

The data employed is challenging in terms of the number of instances and the notorious imbalance in the number of examples per class. To address the first issue, two low dimensional linguistic document representations are proposed, namely the Meta-POS and POS feature characterizations. The second issue is tackled by employing minority over-sampling technique that enables the construction of classifiers from imbalanced datasets. As an empirical instrument to assess the utility of the proposed features, a probabilistic analysis is performed. The results obtained finally show that the set of features presents an acceptable discrimination power in terms of F1 measure in each class.

The proposed representations for the texts transcribed from the oral narratives are contrasted against a standard characterization based on words, i.e. TF-IDF representation. The TF-IDF originally proposed for document retrieval has been successfully used for text categorization across different domains besides its general coverage. Moreover, four classifiers coming from different model families (distance based, quadratic optimization and ensembles) are used over several datasets (generated from the collected data) in order to evaluate how far each representation allows to solve the undertaken task. The results show that TF-IDF allows the highest performance, nevertheless the results attained by using the linguistic features are comparable and also more stable across the different classifiers.

As an overall conclusion, we expect that the proposed features enable a novel and successful approach to identify potential schizophrenic subjects that help out clinic specialists in the early detection of this mental illness. Furthermore, another important issue was the use of a different approach in terms of collecting data, as the data was obtained through an specific linguistic task, which ensure the homogeneity of the data sets. This uniformity contributes to the performance of the models, since it ensures that the narratives are bounded in terms of oral production possibilities, thus making the task more approachable for the classification models.

As a future task, a thorough analysis of the positional dependencies between linguistic features within a text is needed. Additionally, we pose that the gathering of more data will allow to improve substantially the quality of the extracted features and hence the level of discrimination between narrative texts.

REFERENCES

[1] *Diagnostic and Statistical Manual of Mental Disorders*, Vol. 4. American Psychiatric Association, DSM Library, 2014.

[2] R. Tandon, H. A. Nasrallah, and M. S. Keshavan, "Schizophrenia, 'just the facts' 4. Clinical features and conceptualization," *Schizophrenia Res.*, vol. 110, nos. 1–3, pp. 1–23, May 2009.

[3] J. Mcgrath, S. Saha, D. Chant, and J. Welham, "Schizophrenia: A concise overview of incidence, prevalence, and mortality," *Epidemiol. Rev.*, vol. 30, no. 1, pp. 67–76, 2008.

[4] N. C. Andreasen, "Scale for the assessment of thought, language, and communication (TLC)," *Schizophrenia Bull.*, vol. 12, no. 3, pp. 473–482, 1986.

[5] H. Song *et al.*, "Automatic schizophrenic discrimination on fNIRS by using complex brain network analysis and SVM," *BMC Med. Inform. Decision Making*, vol. 17, no. 3, p. 166, Dec. 2017.

[6] B. Elvevag, P. W. Foltz, D. R. Weinberger, and T. E. Goldberg, "Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia," *Schizophrenia Res.*, vol. 93, nos. 1–3, pp. 304–316, 2007.

[7] T. Joachims, "A statistical learning learning model of text classification for support vector machines," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, New York, NY, USA, 2001, pp. 128–136. doi: 10.1145/383952.383974.

[8] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," *Int. J. Med. Inform.*, vol. 77, no. 2, pp. 81–97, Feb. 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1386505606002747

[9] A. Suhasini, S. Palanivel, and V. Ramalingam, "Multimodel decision support system for psychiatry problem," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 4990–4997, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417410011097

[10] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, and D. C. M. Saade, "A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment," *Comput. Biol. Med.*, vol. 51, pp. 140–158, Aug. 2017.

[11] K. H. Nuechterlein, D. M. Barch, J. M. Gold, T. E. Goldberg, M. F. Green, and R. K. Heaton, "Identification of separable cognitive factors in schizophrenia," *Schizophrenia Res.*, vol. 72, no. 1, pp. 29–39, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0920996404003421

[12] D. Greenstein, J. D. Malley, B. Weisinger, L. Clasen, and N. Gogtay, "Using multivariate machine learning methods and structural mri to classify childhood onset schizophrenia and healthy controls," in *Proc. Frontiers Psychiatry*, vol. 3, Jun. 2012, p. 53.

[13] J. Kambeitz *et al.*, "Detecting neuroimaging biomarkers for schizophrenia: A meta-analysis of multivariate pattern recognition studies," *Neuropsychopharmacology*, vol. 40, no. 7, pp. 1742–1751, 2015.

[14] P. Patel, P. Aggarwal, and A. Gupta, "Classification of schizophrenia versus normal subjects using deep learning," in *Proc. 10th Indian Conf. Comput. Vis., Graph. Image Process.*, New York, NY, USA, Dec. 2016, p. 28. doi: 10.1145/3009977.3010050.

[15] R. Boostani, K. Sadatnezhad, and M. Sabeti, "An efficient classifier to diagnose of schizophrenia based on the EEG signals," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6492–6499, 2009.

[16] M. Sabeti, S. D. Katebi, R. Boostani, and G. W. Price, "A new approach for EEG signal classification of schizophrenic and control participants," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2063–2071, 2011.

[17] J. K. Johannesen, J. Bi, R. Jiang, J. G. Kenney, and C.-M. A. Chen, "Machine learning identification of EEG features predicting working memory performance in schizophrenia and healthy adults," *Neuropsychiatric Electrophysiol.*, vol. 2, no. 1, p. 3, 2016.

[18] M. Shim, H.-J. Hwang, D.-W. Kim, S.-H. Lee, and C.-H. Im, "Machine-learning-based diagnosis of schizophrenia using combined sensor-level and source-level EEG features," *Schizophrenia Res.*, vol. 176, nos. 2–3, pp. 314–319, 2016.

[19] U. Grasemann, R. Miikkulainen, and R. Hoffman, "A subsymbolic model of language pathology in schizophrenia," in *Proc. 29th Annu. Conf. Cogn. Sci. Soc.*, Hillsdale, NJ, USA, 2007, pp. 311–316. [Online]. Available: http://nn.cs.utexas.edu/?grasemann:cogsci07

[20] U. Grasemann, R. Miikkulainen, and R. Hoffman, "Modeling acute and compensated language disturbance in schizophrenia," in *Proc. 29th Annu. Conf. Cogn. Sci. Soc.*, 2011, pp. 311–316.

[21] M. Mitchell, K. Hollingshead, and G. Coppersmith, "Quantifying the language of schizophrenia in social media," in *Proc. 2nd Workshop Comput. Linguistics Clin. Psychol., Linguistic Signal Clin. Reality*, Denver, CO, USA, Jun. 2015, pp. 11–20. [Online]. Available: http://www.aclweb.org/anthology/W15-1202

[22] R. D. Strous, M. Koppel, J. Fine, S. Nachliel, G. Shaked, and A. Z. Zivotofsky, "Automated characterization and identification of schizophrenia in writing," *J. Nervous Mental Disease*, vol. 197, no. 8, pp. 585–598, 2009.

[23] 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," *J. Abnormal Psychol.*, vol. 107, no. 3, pp. 423–434, 1998.

[24] G. R. Kuperberg, D. A. Kreher, D. Goff, P. K. McGuire, and A. S. David, "Building up linguistic context in schizophrenia: Evidence from self-paced reading," *Neuropsychology*, vol. 20, 4, pp. 442–452, 2006.

[25] K. McManus, E. Mallory, R. L. Goldfeder, W. A. Haynes, and J. D. Tatum, "Mining Twitter data to improve detection of schizophrenia," in *Proc. AMIA Summits Transl. Sci.*, Mar. 2015, pp. 122–126.

[26] C. Spironelli, A. Angrilli, and L. Stegagno, "Failure of language lateralization in schizophrenia patients: An ERP study on early linguistic components," *J. Psychiatry Neurosci.*, vol. 33, no. 3, pp. 235–243, 2008.

[27] W. Hinzen and J. Rosselló, "The linguistics of schizophrenia: Thought disturbance as language pathology across positive symptoms," in *Proc. Front. Psychol.*, Jul. 2015, p. 971.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[29] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[30] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, 1988.

HÉCTOR ALLENDE-CID received the Ph.D. degree from Universidad Técnica Federico Santa María, Chile, in 2015. He is currently an Assistant Professor with the Escuela de Ingeniería Informática of Pontificia, Universidad Católica de Valparaíso. His research interests include supervised algorithms, distributed regression methods, and image processing.

JUAN ZAMORA received the Ph.D. degree from Universidad Técnica Federico Santa María, Chile, in 2016. He is currently an Assistant Professor with the Instituto de Estadística, Pontificia Universidad Católica de Valparaíso. His research interests include data mining, text mining, and clustering algorithms.

PEDRO ALFARO-FACCIO received the Ph.D. degree from the Pontificia Universidad Católica de Valparaíso, Chile, in 2015, where he is currently an Assistant Professor with the Instituto de Literatura y Ciencias del Lenguaje. His research area consists in studying the Spanish language structure from a psycholinguistics and computational perspective.

MARÍA FRANCISCA ALONSO-SÁNCHEZ received the degree in audiology in Chile and the Ph.D. degree in neuroscience from the Universidad de Zaragoza, Spain. She is currently a Researcher with the Centro de Investigación del Desarrollo en Cognición y Lenguaje, Universidad de Valparaíso, Chile. Her research consists in studying the relationship between language and cognition.