Psychiatry Research 304 (2021) 114135

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Detecting formal thought disorder by deep contextualized word representations

Justyna Sarzynska-Wawer [*,a], Aleksander Wawer [1,b], Aleksandra Pawlak [2,c], Julia Szymanowska [2,c], Izabela Stefaniak [d], Michal Jarkiewicz [3,d], Lukasz Okruszek [a]

[a] Institute of Psychology, Polish Academy of Sciences, Jaracza 1, 00–378 Warszawa, Poland
[b] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01–248 Warszawa, Poland
[c] University of Social Sciences and Humanities, Chodakowska 19/31, 03–815 Warszawa, Poland
[d] Institute of Psychiatry and Neurology, Sobieskiego 9, 02–957 Warszawa, Poland

ARTICLE INFO

Keywords:
Schizophrenia
Language
Natural language processing
Deep learning

ABSTRACT

Computational linguistics has enabled the introduction of objective tools that measure some of the symptoms of schizophrenia, including the coherence of speech associated with formal thought disorder (FTD). Our goal was to investigate whether neural network based utterance embeddings are more accurate in detecting FTD than models based on individual indicators. The present research used a comprehensive Embeddings from Language Models (ELMo) approach to represent interviews with patients suffering from schizophrenia (N=35) and with healthy people (N=35). We compared its results to the approach described by Bedi et al. (2015), referred to here as the coherence model. Evaluations were also performed by a clinician using the Scale for the Assessment of Thought, Language and Communication (TLC). Using all six TLC questions the ELMo obtained an accuracy of 80% in distinguishing patients from healthy people. Previously used coherence models were less accurate at 70%. The classifying clinician was accurate 74% of the time. Our analysis shows that both ELMo and TLC are sensitive to the symptoms of disorganization in patients. In this study methods using text representations from language models were more accurate than those based solely on the assessment of FTD, and can be used as measures of disordered language that complement human clinical ratings.

1. Introduction

Schizophrenia is a severe neuropsychiatric disorder that affects about 1% of the worlds population (Fischer and Buchanan, 2013). It impairs many aspects of functioning including cognitive, affective and social aspects of behavior. Despite the fact that a great deal of effort has been directed towards creating a biopsychosocial model of schizophrenia and utilizing neurobiological markers for its diagnostic process, schizophrenia is still formally diagnosed primarily on the basis of the patient's self-report and clinical observation. Given the marked between-subject heterogeneity of the disorders symptoms, as well as changes in their manifestation in each subject during the course of the disorder, particular stress has been placed on the importance of identifying objective markers of schizophrenia. Formal thought disorder (FTD) may be seen as an example of a symptom of schizophrenia that was recently suggested to be reconceptualized in accordance with Research Domain Criteria (RDoC postulates) (Cohen et al., 2017). Furthermore, studies on neurobiological markers associated with FTD provided insights into its association with structural and functional alterations of language networks in patients (Cavelti et al., 2018). FTD is linked to abnormal speech patterns in patients and may be subdivided into positive (disorganized and incoherent language output) and negative (poverty of speech and language) dimensions (Radanovic et al., 2013). Behavioral studies have emphasized the marked heterogeneity of FTD, with 18 different overall categories of signs being measured by a gold standard tool for the assessment of FTD in patients with

* Corresponding author. .
E-mail addresses: jsarzynska@psych.pan.pl (J. Sarzynska-Wawer), axw@ipipan.waw.pl (A. Wawer), apawlak11@st.swps.edu.pl (A. Pawlak), j.szmnsk@gmail.com (J. Szymanowska), blaszczuk@poczta.onet.pl (I. Stefaniak), michal.mateusz.jarkiewicz@gmail.com (M. Jarkiewicz), lokruszek@psych.pan.pl (L. Okruszek).
[1] (+48) 22 3800500
[2] (+48) 22 5179922
[3] (+48) 22 4582800

https://doi.org/10.1016/j.psychres.2021.114135
Received 4 October 2020; Received in revised form 21 July 2021; Accepted 22 July 2021
Available online 24 July 2021

schizophrenia (Andreasen, 1986). The search for FTD endophenotypes has consistently linked them with impaired executive functioning in patients (Kerns and Berenbaum, 2002; Remberk et al., 2012). Neuroimaging studies have pointed towards an association between FTD and reduced bilateral grey matter volumes in the inferior frontal gyrus, superior temporal gyrus, and the inferior parietal lobe. In addition, resting- and task-related abnormalities in the aforementioned brain structures are also attributed to FTD (Cavelti et al., 2018). Furthermore, neurophysiological studies have observed a relationship between reduced N400 and FTD (Kostova et al., 2005). Altogether these findings support the view that, despite their behavioral heterogeneity, FTD may be closely linked to a specific set of cognitive deficits and neural abnormalities in patients.

FTD is usually diagnosed on the basis of clinical observation of disorganized speech (Adler et al., 1999). Clinicians may also use assessment instruments and scales such as the Thought and Language Disorder (TALD) (Kircher et al., 2014) or Scale for the Assessment of Thought, Language and Communication (TLC), but most of them do not capture all of the elements of FTD or require extensive training to administer. However, recent advances in machine learning and natural language processing have enabled the automatic detection of FTD markers. These methods can provide supplementary information during clinical decision making, as psychiatrists are not always able to notice all of the linguistic differences present in a patient's discourse. It is also difficult to observe subtle changes occurring in the area of thought disorders when following the improvement of a patient's clinical condition after starting treatment. Previous studies using automatic text analysis in schizophrenia focused mostly on decreased speech coherence (Andreasen and Tucker, 1991; Breier and Berg, 1999), which may manifest as a disjointed flow of ideas, a loosening of associations between subsequent words or sequences, or displaying a tangentiality of topics.

Elvevåg et al. (2007) identified a reduction in semantic coherence in patients with schizophrenia, and observed a difference between patients with severe and mild FTD (as assessed by a high global score on the TLC; Andreasen 1986) and healthy controls. In a follow-up study, Elvevåg et al. (2010) examined speech differences between patients with schizophrenia, their first-degree relatives and unrelated healthy people, observing that coherence combined with structural speech analysis accurately differentiates first-degree relatives of patients with schizophrenia from unrelated healthy individuals. This result suggests that even subtle features indicative of underlying genetic vulnerabilities to schizophrenia can be detected with computerized speech analysis. Bedi et al. (2015) measured phrase-level coherence combined with two syntactic markers of speech complexity - maximum phrase length and the use of determiners - and obtained 100% accuracy in predicting later psychosis onset in high-risk youths. Similar results in predicting psychosis were obtained by Corcoran et al. (2018). The most accurate classifier in this study comprised of three different factors among test subjects: decreased semantic coherence, greater variance in that coherence, and reduced use of possessive pronouns.

Studies in which coherence-based models were used showed very promising results. However, this approach does present certain limitations. First of all, it disregards possible FTD markers other than the probability of word combinations within a phrase. FTD might manifest for other linguistic reasons that are unrelated to coherence, such as the tendency to use certain types of words over others, reduction in syntax complexity (Fraser et al., 1986), impaired semantics, and abnormalities in pragmatics (Rodriguez-Ferrera et al., 2001). This is why previous studies on coherence-based models included additional measures such as phrase length or the use of specific parts of speech. However, the number of measures that can be included in the model is limited, and their selection may depend on the language that is being analyzed. In addition, the content of statements may also be important in detecting the linguistic signs of a mental illness (Rezaii et al., 2019), and coherence models only analyze it to a limited extent.

The most recent models that are used in different areas of natural language processing (NLP) address this weakness. They compute word and utterance embeddings. A word (and analogically an utterance) embedding is a learned representation for text where words that have similar meaning have a similar representation. Each word is mapped to one vector, a row of numbers where each point captures some of the word's (utterance's) meaning. For example the vectors for *hamburger* and *cheeseburger* should ideally be very close to each other. The vector for *ferrari* should be far away from those two vectors. Vector values are typically learned in a way that resembles a neural network. Compared to the former class of methods, Embeddings from Language Models (ELMo) (Peters et al., 2018) does not hard-code any assumptions about language features that are usable for detecting FTD. It is a neural network that produces a universally applicable vector representation of utterances. Such vectors are immediately usable not only for FTD but also for tasks such as question answering, or text classification according to various aspects. They contain information about both word order (syntax) and content. ELMo differs from word2vec (see Mikolov et al. (2013a), an algotrithm that learns word embeddings from large bodies of text of text using neural networks) approach by allowing vectors to reflect context-sensitive meaning. For example, when the word book is detected in a sentence it may have different ELMo vector representations depending on neighboring words (as it can be either a verb or a noun, as in *to book a flight* vs *to read a book*).

Our study aimed to examine the ability of such complex and contextual models to detect FTD symptoms, which may be present in patients with schizophrenia. Specifically, we applied ELMo representations with selected supervised classification algorithms. Its accuracy in distinguishing patients with schizophrenia from healthy controls based on written excerpts transcribed from their verbal utterances was compared with the accuracy of classifications obtained from the coherence model and assessments of thinking disorders based on the TLC scale. The coherence model (Bedi et al., 2015) was chosen as the reference point because it is very well recognized in the scientific community (191 citations, as of March 2021). As in this article it uses natural language processing, vectors to represent the meaning of a text, and machine learning for prediction.

Complex models represent information with thousands of features and usually achieve high accuracy, but at the cost of explainability. To address this we decided to use not only clinical and linguistic data, but also to apply Local Interpretable Model-Agnostic Explanations (LIME by Ribeiro et al. (2016); one of the latest models of explainable artificial intelligence) to better understand ELMos predictions. To our knowledge, this is the first study in the field of psychiatry in which such methods have been used.

2. Methods

2.1. Participants

We analyzed statements from thirty-five patients diagnosed with schizophrenia according to ICD-10 criteria (12 females, mean age=35.8, SD=8.53) and a control group consisting of thirty-five demographically matched healthy individuals. Patients who participated in the study were recruited from outpatients being treated at the Institute of Psychiatry and Neurology in Warsaw, Poland. Only clinically stable patients were included in the study, i.e. those who had not undergone any treatment change and had not displayed any significant changes in the severity of their symptoms for a total of four weeks preceding the examination. Patients with intellectual disabilities or a history of comorbid neurological or psychiatric disorders, drug abuse, or any other disorders were also excluded from the study. All of the patients were treated with antipsychotic medication at the time of the study. They had a mean duration of 10.7 years (SD=7.3) of psychiatric treatment with an average of 5 (SD=5) of psychiatric hospitalization. Control group participants were recruited from healthy volunteers who responded to

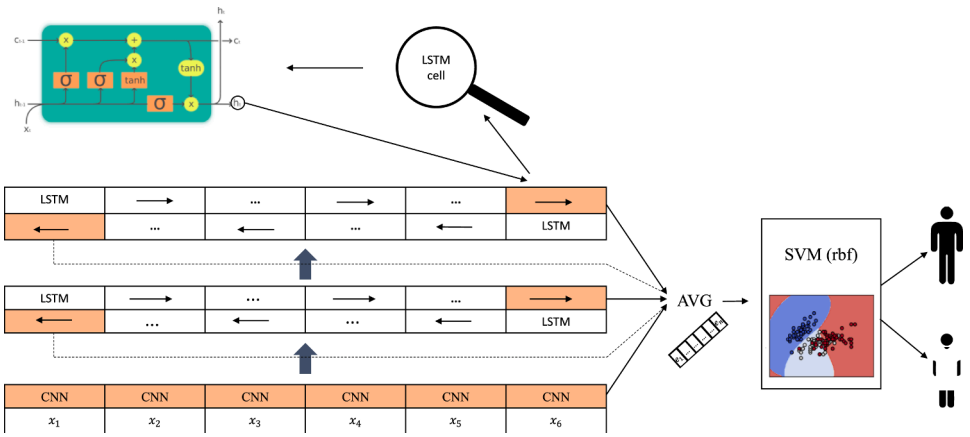

Fig. 1. Detection using ELMo embeddings. The vectors used to create the utterance embeddings are marked in orange: convolutional network (CNN) output layers that represent words and the hidden end states of the long short-term memory (LSTM) neural network of the two layers after reading the entire utterance in each direction. The vectors marked in orange were then averaged (AVG) and used in the classification model SVM (rbf), which divides the statements into those from healthy controls or FTD. The image in green illustrates a single LSTM cell. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

online advertisements. They were paired with patients with schizophrenia based on their sex, age and parental education.

2.2. Clinical interviews

During the clinical interviews with a qualified psychiatrist (MJ), participants underwent Structured Clinical Interviews for the Positive and Negative Syndrome Scale (PANSS) (Kay et al., 1987) and completed the Polish version of the Scale for the Assessment of Thought, Language and Communication (TLC) (Czernikiewicz, 2004). PANSS scores were calculated in line with the five dimension modelvan der Gaag et al. (2006) (Positive: 14.5+/-6.2; Negative: 16.6+/-5.4; Disorganization: 19.7+/-4.7; Excitability: 12.5+/-3.7; Emotional: 16.9+/-6.4). A general TLC score (5.5+/-3.6) was calculated by summing all TLC items.

2.3. Evaluation/classification

We compared three methods used to classify subjects as patients vs. healthy individuals by using two established methods described in the literature (TLC scale, automatized assessment of coherence) and one novel method (ELMo).

2.3.1. TLC Scale

To elicit the sample of speech for the evaluation six questions were asked, four of which respectively concern the patient and their family, the person closest to the patient, and their interests and childhood. Two of the questions are more abstract - they ask why people get sick and why people believe in God. Patients answered these questions as a part of the initial clinical assessment in a hospital, whereas the individuals from the control group were interviewed at the Institute of Psychology of the Polish Academy of Sciences. All subject responses were recorded and then transcribed. Two variants of the transcriptions were prepared: a full version and an edited one (in the case of patients, we removed fragments in which they admitted to be diagnosed with schizophrenia or narrated about the disease, symptoms, hospitalizations, therapies or mental illnesses of parents or other family members; in the statements of people from the control group, we removed the fragments concerning hospitalization due to somatic diseases and concerning mental illnesses in the family). Edited transcripts were 3,2% shorter than their full versions. The edited versions of the transcriptions from both groups were assessed on the TLC Scale by a qualified psychologist. The expert was guided solely by a given statements characteristics, without taking its content into account. In accordance with the TLC scale a psychologist assessed 18 language characteristics, the presence of which was marked on a scale of 0 to 3 or 4 points (depending on the item). These include: poverty of speech, illogicality, incoherence, clanging, neologisms, word approximations, poverty of content of speech, pressure of speech, distractible speech, tangentiality, derailment, stilted speech, echolalia, self reference, circumstantiality, loss of goal, perseveration and blocking.

2.3.2. Coherence model

In order to compute language coherence, we followed the procedure described in Bedi et al. (2015) as closely as possible[4]. We computed two types of coherence features: (1) first-order coherence (FOC) - the similarity of consecutive phrase vectors, averaged over all the phrases in the text, and (2) second-order coherence (SOC) - the similarity between phrases separated by another intervening phrase, averaged over all the phrases in the text. The purpose of these calculations was to identify adjacent but thematically distant phrases. This makes it possible to measure average linguistic consistency and to identify instances of jumping from topic to topic. We calculated the minimum, mean, median, and standard deviation for each type of coherence. Therefore, coherence features contributed 8 variables to each feature vector. Following Bedi et al. (2015) we combined them with features computed from the results of POS (part-of-speech) tagging, using the frequency of each tag as a feature. This contributes 15 feature variables for each text, with the number of features for each text totaling 23. To compare with ELMo, we computed the features separately for each question and concatenated them into one feature vector.

There are three notable differences compared to Bedi et al. (2015), mostly resulting from difference in available resources and tools.

The first one is that instead of LSA, we used Polish language word2vec word embedding vectors to represent words. The LSA algorithm generates a word co-occurrence matrix from a large text corpus, followed by reducing the dimensions of this count matrix using singular value decomposition (SVD). For a more detailed explanation of SVD see Elvevåg et al. (2007). Its popularity is slowly declining as it has been outperformed by more recent algorithms such as word2vec (Mikolov et al., 2013b) or Glove (Pennington et al., 2014) on multiple tasks (Altszyler et al., 2016; Naili et al., 2017). In Bedi et al. (2015), LSA was trained on the Touchstone Applied Science Associates (TASA) Corpus, a collection of educational materials compiled by TASA. No similar

[4] We are aware of multiple studies involving computation of coherence. For example, Hoffman et al. (2018) measured the speech of the elderly. Latent semantic analysis (LSA) was used to calculate the contents of 20 word long windows of text. Coherence was assessed by measuring the similarity of the vector for each window with that of the previous window (local coherence) and with a vector representing the typical semantics of responses to the same prompt (global coherence). LSA and coherence was also used in Iter et al. (2018) for schizophrenia detection. We selected Bedi et al. (2015) as our reference as it was the most cited one and reported 100% accuracy. It is important to note that the model includes not just phrase-level coherence (arguably the most important and novel) but also features such as part-of-speech frequencies.

resource exists for the Polish language, to our knowledge.

In word2vec (Mikolov et al., 2013b), word embedding vectors used to represent words are pulled from a neural network trained using a language modeling task. The Polish word2vec vectors used in our work[5] were trained using a compilation of both Polish Wikipedia and the National Corpus of Polish. Our repository included over two million word vectors, including inflected case-sensitive word forms. We used vectors of 100 elements[6].

The second difference is that we used a dependency parser (a tool for analyzing grammatical structure of sentences) from the Spacy package[7] to obtain phrases, while Bedi et al. (2015) used the NLTK Natural Language Toolkit[8], which to our knowledge does not support the Polish language for phrase extraction.

The third difference is the classification algorithm: a method for assigning a class label to input utterance vectors. Instead of convex hull (see Bedi et al. (2015) for more details) we used the Support Vector Machines (SVM) with radial basis kernel (rbf). SVM classifier seeks a plane that would separate our observations according to their classes, with one class on one side of the plane, while the other class is on the other side. The plane should be as far away from both classes as possible. To help distinguish between classes, the mapping function is used to transform observations. This function (called kernel, radial or rbf is an example) maps data points with lower dimensions into a more dimensional space where they can be more easily separated by linear planes.

We selected this algorithm as it is not only known to perform well in non-linear classification problems, but also strongly established in the context of text classification based on embedding vectors, as evidenced by multiple papers since 2018 (Hettinger et al., 2018; Indurthi et al., 2019; Lai et al., 2018).

2.3.3. ELMo

We also applied the Embeddings from Language Models (Peters et al., 2018) method to detect FTD from textual utterances. ELMo is a pre-trained, multi-layer, bi-directional neural network language model. For each text input, ELMo uses a convolutional neural network layer (CNN) to generate word embedding vectors, an alternative approach to word2vec. Then, it uses two layers of bidirectional LSTM recurrent neural networks (Hochreiter and Schmidhuber, 1997). Both CNN and LSTM layers are pre-trained on large bodies of text in language modeling tasks.

LSTMs are recurrent neural networks that read word for word using word embedding vectors, and in doing so they remember words seen in the past. Memorization is accomplished by the cell's internal vector (also called state, hidden state, or memory). When an LSTM cell reads the input word vector, it updates the memory vector. The content of the memory vector (hidden state) after reading the entire text can be used as a representation of its meaning[9]

Text representations are produced using ELMo by obtaining the averaging of three layers: one CNN layer and two LSTM layers. In the case of the LSTM layers, information is extracted from hidden states of LSTM cells after reading the whole input sequence, forward or backward.

Our results were obtained using the Polish language ELMo version (Che et al., 2018) (ELMoForManyLangs). Pre-training was performed on a 20-million-words data set sampled from Wikipedia and Common Crawl.

For each input utterance, the representation obtained from ELMo is a vector of 1024 elements. The vectors were then classified with the SVM (rbf). SVM was similarly used in the coherence model, allowing comparison between these two methods.

Fig. 1 demonstrates the process of discriminating healthy individuals from individuals with schizophrenia by textual utterances using ELMo representation and supervised classification.

2.4. LIME Model explanations

In order to better understand where the labels calculated by the models come from, we used an algorithm that explains their decisions. Local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016) treats the model as a black box, meaning that it does not attempt to explore its interior and weights, but instead focuses on approximating the model's behavior. LIME only requires a classifier that takes in raw text and outputs a probability for each class. In order to figure out what parts of the interpretable input are contributing to the prediction, LIME perturbs the input and checks how the model's predictions change due to this disruption. Even if the classifier uses some uninterpretable representation such as word embeddings, this method still works. A popular example concerns a sentiment analysis task: in order to explain the sentiment prediction for the sentence "I hate this movie", LIME would compute predictions on similar sentences such as "I hate movie", "I this movie", "I movie", "I hate", and so on, possibly leading to the identification of the word hate as relevant. A similar procedure was applied to identify words relevant for classifying utterances as originating from diagnosed vs healthy controls.

3. Results

3.1. Classifications

3.1.1. TLC

TLC scale scores were obtained by summing up points from all of the listed items, and this was used as the criteria for assigning a given person to either a patient group or a control group. The most accurate results were obtained by calculating a median for grouped data. The accuracy of TLC is 74%.

3.1.2. Coherence

Table 1 contains mean accuracy values computed using the coherence method described in Section 2.3.2 in a leave-one-out cross-validation using SVM (rbf). In the leave-one-out cross-validation, the learning algorithm is applied once for each instance, using all other instances as a training set and using the selected instance as a single-item test set. The test results are combined (averaged) over the rounds to give an estimate of the model's predictive performance.

Performance on all questions (1–6) and the best combination is reported along with corresponding mean accuracy and standard error of the cross-validation error estimate (± number). To investigate upper limits of possible accuracy we were looking for the best combination of questions by testing all possible combinations (permutations) of the six interview questions. However, as a reference, we suggest using the more conservative estimates of the model computed for all questions. The best combination should be interpreted with caution.

We computed the experiments on all transcribed utterances without any modifications (the full variant) since mentioning a specific topic (in

Table 1
Coherence model as in Bedi et al. (2015): average accuracy in leave-one-out cross-validation.

Classifier	Questions	Accuracy
SVM (rbf)	1–6 (all)	70% (± 5%)
	1,2,4,6 (best)	75% (± 5%)

[5] http://dsmodels.nlp.ipipan.waw.pl/
[6] File name: nkjp+wiki-forms-all-100-cbow-hs
[7] https://spacy.io/models/pl
[8] https://www.nltk.org
[9] Using this method, one can calculate the utterance embedding vector for any sentence. For example, *The quick brown fox jumps over the lazy dog* might be converted to some row of numbers (vector) such as [0.41, -0.05, ..., 0.98].

Table 2
ELMo: Average accuracy in leave-one-out cross-validation .

Classifier	Questions	Accuracy
SVM (rbf)	1–6 (all)	80% (± 5%)
	2–6 (best)	84% (± 5%)

our case - diseases) has no impact on overall lexical coherence scores.

3.1.3. ELMo

Table 2 contains mean accuracy values computed on the full data variant from ELMo models in leave-one-out cross-validation using the SVM (rbf). Performance on all questions, the best performing combination of questions, and the corresponding mean accuracy and standard error of the cross-validation error estimate (± number) are all reported.

The ELMo representation contains information about both the content and style of utterances, and as such we experimented with two data variants. The full one contains all transcribed utterances. The edited variant contains utterances without text fragments mentioning patients discussing their diagnoses. Such fragments could provide clues and make the task of diagnosing significantly easier for both humans and selected automated methods such as ELMo. Contrary to expectation, the accuracy of the SVM (rbf) was exactly the same for both data types.

3.2. Analysis

ELMo embeddings turned out to be a powerful text representation for detecting FTD. The SVM (rbf) algorithm reached an accuracy of 84% on the optimized feature space. It performed best on questions 2 to 6. The ELMo model was the most accurate, and was able to correctly classify 80% (28 out of 35) of patients and 88% (31 out of 35) of healthy individuals, averaging out to an overall cross-validated classification accuracy of 84%. The accuracy on all six questions reached 80%.

Using the coherence feature set, the SVM (rbf) obtained an overall cross-validated classification accuracy on all six questions as high as 70%. This model correctly classified 26 out of 35 patients (74%) and 23 out of 35 healthy individuals (65%).

Ratings based on the TLC criteria divided by the median had a hit rate of 74% (24 out of 35 recognized patients and 28 out of 35 healthy controls). Cochran's Q test indicated that there are no significant differences in accuracy between the three methods: $\chi^2(2) = 4.79$ ($p = 0.09$). A pairwise post-hoc Durbin-Conover test was significant for ELMo vs. Coherence model ($p = 0.03$).

Table 3 illustrates the errors of three schizophrenia detection methods, as well as sensitivity and specificity.

Table 3
Errors of three schizophrenia detection methods. TP: true positives, TN: true negatives, FP: false positives (type I errors), FN: false negatives (type II errors), Sens: sensitivity, Spec: specificity.

	TP	TN	FP	FN	Sens	Spec
TLC	24	28	7	11	0.68	0.80
ELMo+SVM (rbf)	28	31	4	7	0.80	0.88
Coherence+SVM (rbf)	26	23	12	9	0.74	0.65

3.3. LIME Model explanations

Conclusions drawn from LIME are typically qualitative rather than quantitative. They are based on two mutually related types of information that LIME provides to facilitate model explanations: a set of weights assigned to words, indicating their positive or negative contribution to the model decision, and a text with highlighted words. Colors are linked to each of the classes, and saturation levels to weight values.

In our study, we focused on a small group of people misclassified by the ELMo. The first observation was that weights assigned to the same words differ between utterances. This strongly indicates that ELMo utterance vectors are sensitive to context and word meaning.

The conclusions one can infer from LIME explanations are as follows:

- Frequently repeated words containing little content (e.g. 'somehow', 'some', 'such', 'a bit', 'well') are indicative for FTD classification by the ELMo. This finding is aligned with the role of low semantic density reported by Rezaii et al. (2019).
- Spiritually linked words are sometimes indicative for FTD.
- Work and professional vocabulary are in some of the texts indicative for healthy control classification.
- Family references are usually important, but interestingly play a mixed role. In most cases they are indicative for healthy controls, but multiple opposite examples can also be found.

The examples below are excerpts of actual statements translated from Polish, with important words marked by the LIME explainer. Words highlighted in red contribute to the TFD class, Words highlighted in green to the healthy control decision.

.. Well - mom. I honestly owe her a lot. Well, I feel she loves me very much. [..] Well, my mother and I get on well.

.. And I also think that a person needs a higher authority, a force majeure that is there, a person to whom one can turn for help.

3.4. Sensitivity to symptoms

To examine each method's sensitivity to symptoms, we performed an additional series of Logistic Regressions with five PANSS factors and the outcomes of each type of classification in the patients' group (ELMo, TLC scale, coherence). For the ELMo score, PANSS disorganization accurately classified 86% of ELMo labels [4 FP/ 1 FN]. Similarly, the level of PANSS disorganization symptoms accurately classified 77% of TLC labels. None of the PANSS dimensions significantly predicted coherence-based classification. Patients correctly classified by ELMo presented a significantly higher level of disorganization compared to patients classified as healthy individuals (TP: 20.9+/-4.2 vs. FN: 14.7+/-2.9; $t(33)=3.6$). Similar, albeit less notable, results were observed for TLC classification (TP: 20.9+/-3.8 vs. FN: 16.8+/-5.3 $t(33)=2.6$). Furthermore, secondary analyses of the between-group differences in the severity of all 36 PANSS items showed that misclassified patients present a lower level of Conceptual disorganization (P2; ELMo and TLC) and

Disorientation (G10; ELMo). An analysis of part-of-speech and sentiment in statements was also conducted. The probability of being classified as a patient by ELMo (pulled from model class probabilities) was negatively correlated with positive sentiment ($r = -0.33$; $p=0.005$) and finite verbs ($r = -0.23$; $p=0.05$).

4. Discussion

Is it possible for a computer model trained in less than an hour to become attuned to symptoms of formal thought disorder in patients with schizophrenia? Previous research already established that computer models which distinguish patients from healthy people on the basis of their statements can perform such tasks very effectively. However, NLP is developing at a rapid speed, and we therefore checked if the new complex models could alone perform this task as accurately as diverse feature-based models.

The results we obtained clearly indicate that the ELMo model we used was more accurate than the coherence calculation model, which has been widely used in previous studies within the field (Bedi et al., 2015; Elvevåg et al., 2010; 2007). Furthermore, ELMo achieved a higher accuracy than clinically-based assessment, which was based on a well-known scale for evaluating FTD in schizophrenic patients.

Most studies of automatic FTD recognition are cases of feature engineering. This approach is characteristic of mainstream machine learning from previous decades, and starts with the observation of which linguistic dimension is relevant to FTD. Its obvious limitation is human perceptiveness, as researchers are unlikely to recognize every important phenomenon. Some studies focus on language coherence (Bedi et al., 2015) while others focus on low semantic density (Rezaii et al., 2019). No study to this point has covered both, and it is also very likely that there are more FTD-relevant features. Utterance embedding vectors from deep neural networks (ELMo is just one case of many) have the chance to implicitly capture language characteristics relevant to FTD, including the two mentioned above and many others. ELMo does not hardcode any knowledge or assumptions about what information, including language-dependent features, is important for FTD recognition.

ELMo representation has another interesting property. ELMo utterance vectors, for any language, capture both syntactic and semantic information of the same kind. This is not always the case with feature engineering. Determiners are a good example of feature engineering using a language-specific feature. The model by (Bedi et al., 2015) includes both coherence features - which are possible to replicate in a reasonable way in most languages (including Polish) - and the counts of determiners. Unfortunately, not all languages have a lexically distinct class of determiners. This is true for Polish, where no such class of words is clearly defined. Considering the importance of replicating scientific research of this type in the multilingual context, the advantage of ELMo is clear.

The accuracy obtained by such complex models cause them to already be widely tested for diagnosis of various diseases and conditions. However, they are not flawless. Although they take into account a wide variety of statement characteristics, we are as of yet unable to identify which particular aspects of language are disordered. They themselves don't shed any new light on the mechanisms of schizophrenia or FTD, which indicates it is important to correlate them with clinical ratings and other language indicators.

Recently there are also more explainable AI methods, and it seems likely that thanks to their use we will be able to understand the decisions made by classifiers that have so far been treated as black boxes. We showed a sample of the capabilities of such models on the example of LIME, which revealed that it is possible to identify certain types of words as indicative for either FTD or healthy controls (frequently repeated low content words, spiritual words, professional words and family references). The importance of a given word for FTD or a healthy class depends on the context in which it appears and varies between utterances.

We showed a sample of the capabilities of such models using the example of LIME which revealed that it is possible to identify certain types of words as indicative of FTD or healthy controls (eg, low content frequently repeated words, spiritual words, professional words and family references).

In our study, both ELMo and TLC were sensitive to the symptoms of disorganization (but not other groups of symptoms) in patients: a higher level of disorganization observed during the clinical interview was linked to how accurately a test subject was classified as either a patient or healthy individual. These results suggest that the ELMo model has shown both sensitivity and specificity to disorganization symptoms while classifying patients vs. controls.

According to Foltz et al. (2016), one of the biggest problem in research on automatic text analysis when detecting clinically significant results are the typically small research samples and the need to adapt techniques taken from NLP to our discipline. In our study, we also faced the problem of only gathering a small amount of statements from patients with schizophrenia. For this reason, we used the same data for feature selection (seeking out the best performing combinations of questions) and to compute performance estimates. As a general rule, feature selection should be performed on a separate data set (dev set) than the measurement of accuracy (test set). Unfortunately, separate dev and test sets are not feasible in our case due to insufficient sample size. Therefore, we also report the results on all of the features (questions 1–6) without any form of feature selection. It should be noted, however, that a group of 35 patients with schizophrenia exceeds the sample described in most of the previous studies (e.g. Elvevåg et al. (2010)), especially since each participant gave six statements. Another limitation in our study is an unbalanced sex distribution (12 Female/ 23 Male). There are many linguistic differences between the sexes (Newman et al., 2008) and they may affect the results. However, schizophrenia also occurs more often in men in the population. Depending on the studies, it is 1.4 (McGrath et al., 2004) to 7.5 (Scully et al., 2002) times more common in men than in women. Above all, however, Foltz et al. (2016) emphasize that research on text analysis in patients has been "streamlined": new models (such as ELMo, as was used here) do not require significant assimilation into the field to be used effectively. For this reason, implementing these technologies into clinical encounters may prove to be useful not only in aiding with diagnosis, but also further tool development. Finally, it is worth mentioning that the effectiveness of these models may enable psychiatrists and other specialists in clinical trials to observe subtle changes in patients' speech, which may be useful in the study of FTD pathogenesis or in monitoring patients at ultra-high risk of developing psychosis. In addition, this method has the added benefit of being able to be conducted remotely without the need for a face-to-face interview. These technologies are widely accessible and therefore should be implemented into clinical practice as an additional diagnosis resource. Future studies should focus on developing training on ELMo method for psychiatrists and creating a system in which use of these technologies can also simultaneously result in data collection and improvement of these tools.

CRediT authorship contribution statement

Justyna Sarzynska-Wawer: Conceptualization, Project administration, Visualization, Data curation, Writing – original draft, Writing – review & editing. **Aleksander Wawer:** Software, Formal analysis, Writing – original draft, Writing – review & editing. **Aleksandra Pawlak:** Investigation, Visualization. **Julia Szymanowska:** Investigation, Data curation. **Izabela Stefaniak:** Investigation, Writing – review & editing. **Michal Jarkiewicz:** Resources, Writing – review & editing. **Lukasz Okruszek:** Formal analysis, Writing – original draft, Writing – review & editing, Funding acquisition.

Acknowledgment

Research was funded by The National Science Centre (Poland), grant number UMO-2016/23/D/HS6/02947 awarded to Lukasz Okruszek.

References

Adler, C.M., Malhotra, A.K., Elman, I., Goldberg, T., Egan, M., Pickar, D., Breier, A., 1999. Comparison of ketamine-induced thought disorder in healthy volunteers and thought disorder in schizophrenia. American Journal of Psychiatry 156 (10), 1646–1649.

Altszyler, E., Sigman, M., Slezak, D.F., 2016. Comparative study of LSA vs word2vec embeddings in small corpora: a case study in dreams database. CoRR abs/1610.01520.

Andreasen, N.C., 1986. Scale for the assessment of thought, language, and communication (tlc). Schizophr Bull 12 (3), 473.

Andreasen, N.C., Tucker, G.J., 1991. Introductory textbook of psychiatry. American Journal of Psychiatry 148 (5).670–670

Bedi, G., Carrillo, F., Cecchi, G.A., Slezak, D.F., Sigman, M., Mota, N.B., Ribeiro, S., Javitt, D.C., Copelli, M., Corcoran, C.M., 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr 1, 15030.

Breier, A., Berg, P.H., 1999. The psychosis of schizophrenia: prevalence, response to atypical antipsychotics, and prediction of outcome. Biol. Psychiatry 46 (3), 361–364.

Cavelti, M., Kircher, T., Nagels, A., Strik, W., Homan, P., 2018. Is formal thought disorder in schizophrenia related to structural and functional aberrations in the language network? a systematic review of neuroimaging findings. Schizophr. Res. 199, 2–16.

Che, W., Liu, Y., Wang, Y., Zheng, B., Liu, T., 2018. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics, Brussels, Belgium, pp. 55–64. http://www.aclweb.org/anthology/K18-2005.

Cohen, A.S., Le, T.P., Fedechko, T.L., Elvevåg, B., 2017. Can rdoc help find order in thought disorder? Schizophr Bull 43 (3), 503–508.

Corcoran, C.M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D.C., Bearden, C.E., Cecchi, G.A., 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. World Psychiatry 17 (1), 67–75.

Czernikiewicz, A., 2004. Przewodnik po zaburzeniach językowych w schizofrenii. Instytut Psychiatrii i Neurologii.

Elvevåg, B., Foltz, P.W., Rosenstein, M., DeLisi, L.E., 2010. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. J Neurolinguistics 23 (3), 270–284.

Elvevåg, B., Foltz, P.W., Weinberger, D.R., Goldberg, T.E., 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophr. Res. 93 (1–3), 304–316.

Fischer, B.A., Buchanan, R.W., 2013. Schizophrenia in adults: epidemiology and pathogenesis. U: UpToDate, Post TW ur. UpToDate [Internet]. Waltham, MA: UpToDate.

Foltz, P.W., Rosenstein, M., Elvevåg, B., 2016. Detecting clinically significant events through automated language analysis: quo imus? NPJ Schizophr 2, 15054.

Fraser, W.I., King, K.M., Thomas, P., Kendell, R.E., 1986. The diagnosis of schizophrenia by language analysis. The British Journal of Psychiatry 148 (3), 275–278.

van der Gaag, M., Hoffman, T., Remijsen, M., Hijman, R., de Haan, L., van Meijel, B., van Harten, P.N., Valmaggia, L., De Hert, M., Cuijpers, A., et al., 2006. The five-factor model of the positive and negative syndrome scale ii: a ten-fold cross-validation of a revised model. Schizophr. Res. 85 (1–3), 280–287.

Hettinger, L., Dallmann, A., Zehe, A., Niebler, T., Hotho, A., 2018. ClaiRE at SemEval-2018 task 7: Classification of relations using embeddings. Proceedings of The 12th International Workshop on Semantic Evaluation. Association for Computational Linguistics, New Orleans, Louisiana, pp. 836–841. https://doi.org/10.18653/v1/S18-1134. https://www.aclweb.org/anthology/S18-1134.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput 9 (8), 1735–1780.

Hoffman, P., Loginova, E., Russell, A., 2018. Poor coherence in older people's speech is explained by impaired semantic and executive processes. Elife 7. https://doi.org/10.7554/elife.38907.

Indurthi, V., Syed, B., Shrivastava, M., Chakravartula, N., Gupta, M., Varma, V., 2019. FERMI at SemEval-2019 task 5: Using sentence embeddings to identify hate speech against immigrants and women in Twitter. Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Minneapolis, Minnesota, USA, pp. 70–74. https://doi.org/10.18653/v1/S19-2009. https://www.aclweb.org/anthology/S19-2009.

Iter, D., Yoon, J., Jurafsky, D., 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. Association for Computational Linguistics, New Orleans, LA, pp. 136–146. https://doi.org/10.18653/v1/W18-0615. https://www.aclweb.org/anthology/W18-0615.

Kay, S.R., Fiszbein, A., Opler, L.A., 1987. The positive and negative syndrome scale (panss) for schizophrenia. Schizophr Bull 13 (2), 261–276.

Kerns, J.G., Berenbaum, H., 2002. Cognitive impairments associated with formal thought disorder in people with schizophrenia. J Abnorm Psychol 111 (2), 211.

Kircher, T., Krug, A., Stratmann, M., Ghazi, S., Schales, C., Frauenheim, M., Turner, L., Fährmann, P., Hornig, T., Katzev, M., et al., 2014. A rating scale for the assessment of objective and subjective formal thought and language disorder (tald). Schizophr. Res. 160 (1–3), 216–221.

Kostova, M., Passerieux, C., Laurent, J.-P., Hardy-Baylé, M.-C., 2005. N400 Anomalies in schizophrenia are correlated with the severity of formal thought disorder. Schizophr. Res. 78 (2–3), 285–291.

Lai, S., Leung, K.S., Leung, Y., 2018. SUNNYNLP at SemEval-2018 task 10: A support-vector-machine-based method for detecting semantic difference using taxonomy and word embedding features. Proceedings of The 12th International Workshop on Semantic Evaluation. Association for Computational Linguistics, New Orleans, Louisiana, pp. 741–746. https://doi.org/10.18653/v1/S18-1118. https://www.aclweb.org/anthology/S18-1118.

McGrath, J., Saha, S., Welham, J., El Saadi, O., MacCauley, C., Chant, D., 2004. A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. BMC Med 2 (1), 13.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed Representations of Words and Phrases and Their Compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26. Curran Associates, Inc., pp. 3111–3119

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed Representations of Words and Phrases and Their Compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 26. Curran Associates, Inc., pp. 3111–3119

Naili, M., Habacha, A., Ben Ghezala, H., 2017. Comparative study of word embedding methods in topic segmentation. Procedia Comput Sci 112, 340–349.

Newman, M.L., Groom, C.J., Handelman, L.D., Pennebaker, J.W., 2008. Gender differences in language use: an analysis of 14,000 text samples. Discourse Process 45 (3), 211–236.

Pennington, J., Socher, R., Manning, C., 2014. GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162. https://www.aclweb.org/anthology/D14-1162.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Association for Computational Linguistics, pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202.

Radanovic, M., Sousa, R.T.d., Valiengo, L., Gattaz, W.F., Forlenza, O.V., 2013. Formal thought disorder and language impairment in schizophrenia. Arq Neuropsiquiatr 71 (1), 55–60.

Remberk, B., Namysłowska, I., Rybakowski, F., 2012. Cognition and communication dysfunctions in early-onset schizophrenia: effect of risperidone. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 39 (2), 348–354.

Rezaii, N., Walker, E., Wolff, P., 2019. A machine learning approach to predicting psychosis using semantic density and latent content analysis. NPJ Schizophr 5 (1), 1–12.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "why should i trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, NY, USA, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.

Rodriguez-Ferrera, S., McCarthy, R., McKenna, P., 2001. Language in schizophrenia and its relationship to formal thought disorder. Psychol Med 31 (2), 197–205.

Scully, P.J., Quinn, J.F., Morgan, M.G., Kinsella, A., O'Callaghan, E., Owens, J.M., Waddington, J.L., 2002. First-episode schizophrenia, bipolar disorder and other psychoses in a rural irish catchment area: incidence and gender in the cavan–monaghan study at 5 years. The British Journal of Psychiatry 181 (S43), s3–s9.