TecnoLógicas
ISSN-p 0123-7799
ISSN-e 2256-5337
Vol. 23, No. 47, pp. 63-75
Enero-abril de 2020

Artículo de Revisión/Research Article

Word-Embeddings and Grammar Features to Detect Language Disorders in Alzheimer's Disease Patients

Detección de desórdenes de lenguaje de pacientes con enfermedad de Alzheimer usando embebimientos de palabras y características gramaticales

Juan S. Guerrero-Cristancho [1],
Juan C. Vásquez-Correa [2], y
Juan R. Orozco-Arroyave [3]

Recibido: 13 de junio de 2018
Aceptado: 3 de octubre de 2019

Cómo citar / How to cite
J. S. Guerrero-Cristancho, J. C. Vásquez-Correa, J. R. Orozco-Arroyave, "Word-Embeddings and Grammar Features to Detect Language Disorders in Alzheimer's Disease Patients," *TecnoLógicas*, vol. 23, no. 47, pp. 63-75, 2020. https://doi.org/10.22430/22565337.1387

[1] Estudiante de Ingeniería Electrónica, Grupo de investigación en Telecomunicaciones aplicadas (GITA), Facultad de Ingeniería, Universidad de Antioquia, Medellín-Colombia, jsebastian.guerrero@udea.edu.co
[2] MSc. en Ingeniería de Telecomunicaciones, Grupo de investigación en Telecomunicaciones aplicadas (GITA), Facultad de Ingeniería, Universidad de Antioquia, Laboratorio de reconocimiento de patrones (LME), Universidad de Erlangen, Erlangen-Alemania. jcamilo.vasquez@udea.edu.co
[3] PhD. en Ciencias de la Computación, Grupo de investigación en Telecomunicaciones aplicadas (GITA), Facultad de Ingeniería, Universidad de Antioquia, Laboratorio de reconocimiento de patrones (LME), Universidad de Erlangen, Erlangen-Alemania, rafael.orozco@udea.edu.co

Abstract

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that affects the language production and thinking capabilities of patients. The integrity of the brain is destroyed over time by interruptions in the interactions between neuron cells and associated cells required for normal brain functioning. AD comprises deterioration of the communicative skills, which is reflected in deficient speech that usually contains no coherent information, low density of ideas, and poor grammar. Additionally, patients exhibit difficulties to find appropriate words to structure sentences. Multiple ongoing studies aim to detect the disease considering the deterioration of language production in AD patients. Natural Language Processing techniques are employed to detect patterns that can be used to recognize the language impairments of patients. This paper covers advances in pattern recognition with the use of word-embedding and word-frequency features and a new approach with grammar features. We processed transcripts of 98 AD patients and 98 healthy controls in the Pitt Corpus of the Dementia-Bank database. A total of 1200 word-embedding features, 1408 Term Frequency—Inverse Document Frequency features, and 8 grammar features were extracted from the selected transcripts. Three models are proposed based on the separate extraction of such feature sets, and a fourth model is based on an early fusion strategy of the proposed feature sets. All the models were optimized following a Leave-One-Out cross validation strategy. Accuracies of up to 81.7 % were achieved using the early fusion of the three feature sets. Furthermore, we found that, with a small set of grammar features, accuracy values of up to 72.8 % were obtained. The results show that such features are suitable to effectively classify AD patients and healthy controls.

Keywords

Alzheimer's Disease, Natural Language Processing, Text Mining, Classification, Machine Learning.

Resumen

La enfermedad de Alzheimer es un desorden neurodegenerativo-progresivo que afecta la producción de lenguaje y las capacidades de pensamiento de los pacientes. La integridad del cerebro es destruida con el paso del tiempo por interrupciones en las interacciones entre neuronas y células, requeridas para su funcionamiento normal. La enfermedad incluye el deterioro de habilidades comunicativas por un habla deficiente, que usualmente contiene información inservible, baja densidad de ideas y habilidades gramaticales. Adicionalmente, los pacientes presentan dificultades para encontrar palabras apropiadas y así estructurar oraciones. Por lo anterior, hay investigaciones en curso que buscan detectar la enfermedad considerando el deterioro de la producción de lenguaje. Así mismo, se están usando técnicas de procesamiento de lenguaje natural para detectar patrones y reconocer las discapacidades del lenguaje de los pacientes. Por su parte, este artículo se enfoca en el uso de características basadas en embebimiento y frecuencia de palabras, además de hacer una nueva aproximación con características gramaticales para clasificar la enfermedad de Alzheimer. Para ello, se consideraron transcripciones de 98 pacientes con Alzheimer y 98 controles sanos del Pitt Corpus incluido en la base de datos Dementia-Bank. Un total de 1200 características de embebimientos de palabras, 1408 características de frecuencia de término inverso vs. frecuencia en documentos, y 8 características gramaticales fueron calculadas. Tres modelos fueron propuestos, basados en la extracción de dichos conjuntos de características por separado y un cuarto modelo fue basado en una estrategia de fusión temprana de los tres conjuntos de características. Los modelos fueron optimizados usando la estrategia de validación cruzada Leave-One-Out. Se alcanzaron tasas de aciertos de hasta 81.7 % usando la fusión temprana de todas las características. Además, se encontró que un pequeño conjunto de características gramaticales logró una tasa de acierto del 72.8 %. Así, los resultados indican que estas características son adecuadas para clasificar de manera efectiva entre pacientes de Alzheimer y controles sanos.

Palabras clave

Enfermedad de Alzheimer, procesamiento de lenguaje natural, minería de texto, clasificación, aprendizaje de máquina.

1. INTRODUCTION

Alzheimer's Disease (AD) is the most common type of neurodegenerative dementia; it disturbs the interactions between neuron cells involved in the brain functions, isolating them [1], [2].

Communication and cognitive skills are affected in AD patients [3]. For instance, the production of language, coherent sentences, and capabilities to structure conversations are compromised [4], [5].

Language production, both spoken and written, shows deficient speech that usually contains a high number of words and verbal utterances with no coherent information, low density of ideas, and poor grammar [6], [7]. Conversation structuring is undermined by the scarcity of declarative sentences such as propositions. Additionally, patients use pronouns more frequently and have a hard time finding the right words for a sentence [8].

The process to diagnose AD is difficult and time-consuming. However, speech and language processing can help, and its first step is the automatic classification of AD patients and healthy controls (HC). For that reason, the interest of the research community in contributing to the AD detection process has increased in recent years.

In [9], the authors classified transcripts from 99 AD patients and 99 HC subjects from the Dementia-Bank dataset [10].

They extracted syntactic, lexical, and n-gram-based features for the classification [11]. The syntactic features included the number of occurrences of coordinated, subordinated, and reduced sentences per patient, number of predicates, and average number of predicates. The lexical features comprised the total number of utterances, average length of utterances, and number of unique words and function words, among others. The features were classified using a Support Vector Machine (SVM). The models were validated using a Leave-Pair-Out-Cross-Validation strategy.

They reported accuracy values of up to 93 % with the proposed features. The same database was used in [12], where the authors classified AD and HC participants using a model based on word embeddings. The word-embedding technique they used was based on the Global Vectors (GloVe) model, which considers the context of neighbor words and the word occurrence in a document [13]. Said authors considered a pre-trained model with the Common Crawl dataset, whose vocabulary size exceeds the 2 million and contains 840 billion words.

A logistic regression classifier and a Convolutional Neural Network with Long Short-Term Memory Units (CNN-LSTM) were implemented for the classification. The models were validated with a 10-Fold-Cross-Validation strategy, and the authors reported accuracies of up to 75.6 %.

In [14], AD patients in the Dementia-Bank dataset were classified using a Bag-of-Words (BoW) representation and a classifier based on neural networks. The parameters of the classifier were optimized following a Leave-One-Out cross validation strategy (LOO), and an accuracy of 91 % was reported. In [15], the authors used a hybrid model composed of Word2Vec (W2V) word-embeddings, Term Frequency—Inverse Document Frequency (TF-IDF) features, and Latent Dirichlet Allocation (LDA) topic probabilities [16], [17], [18]. In that case, the Dementia-Bank dataset was used along with the 2011 survey of the Wisconsin Longitudinal Study (WLS) [19].

They considered an SVM classifier with a linear kernel whose complexity parameter was optimized following a 5-Fold-Cross-Validation strategy. The authors reported an accuracy of 77.5 % and established that the most accurate features were those based on TF-IDF combined with the W2V model.

This study considers word-embedding features extracted from a W2V model trained with the latest data dump from Wikipedia (February 2019), along with TF-

IDF features and grammatical features, in order to classify AD patients and HC subjects based on transcripts in the Dementia-Bank dataset. The results show that the W2V model, along with an early fusion of the three feature sets, is appropriate to model the cognitive impairments of the patients. Additionally, the results indicate that the grammatical features are suitable to identify HC subjects and AD patients. To the best of our knowledge, this is the first study that considers grammatical features to model language deficiencies exhibited by AD patients.

2. METHODS

2.1 Word Embeddings

The global coherence of the spontaneous speech of AD patients shows semantic, comprehension, and memory loss impairments. Semantic impairments include errors when naming objects or actions [6]. Contextual impairments result in incorrect categorical names for entities and incoherent information in sentences [20].

Memory impairments are reflected in the restricted vocabulary of AD patients and their difficulties to find appropriate words for sentences [6]. W2V considers the contextual relations between words and their co-occurrences in a transcript. In this study, we aim to detect the impairments mentioned above in AD patients using word embeddings extracted from a W2V model.

The words in the selected transcripts of the dataset are mapped into vectors that are positioned in a n-dimensional space according to their context. On the one hand, the closer the word vectors, the more related the words are in that context. On the other hand, the further the vectors are from each other, the less the words are related in that context. Such relationships are illustrated in Fig. 1.

W2V learns from the co-occurrence information of words and is based on two architectures: Skipgram and Continuous Bag-of-words (CBoW). This study implemented the CBoW architecture, which is designed to predict a word for a given context. The W2V model is based on a neural network with a single hidden layer. This architecture is trained with examples from a given context in order to predict a word in the output [16].

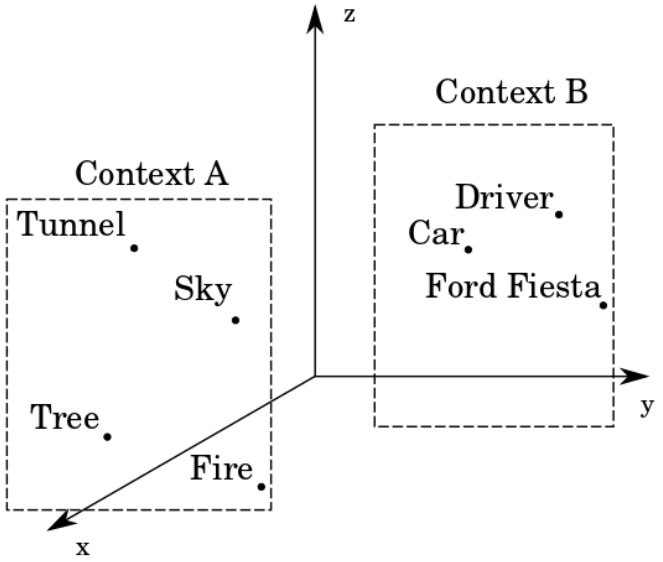

Fig. 1. Illustrative example of word embeddings. Source: Created by the authors.

The vocabulary size is v. The Y words of the context $\{Y_0, Y_1, \dots, Y_c\}$ are the one-hot encoded inputs of the $\{X_0, X_1, \dots, X_v\}$ neurons at the input layer. The hidden layer has h_n neurons, where n is the dimension of the W2V model. The output layer has O_v neurons. The values $\{O_0, O_1, \dots, O_v\}$ are used to predict the most probable word for the input context word W. This process is shown in Fig. 2.

Said model was trained with the latest Wikipedia data dump (February 2019).

The vocabulary size of the model is over 2 million and it has over 2 billion words. The Gensim topic modeling toolkit was used to develop the W2V model [21]. Default parameters were used unless specified. The feature extraction process consists of four steps: 1) The stop words are removed from the documents using the English stop words dictionary available in the Natural Language Toolkit (NLTK) [22]. 2) The W2V model is trained with the processed text corpus, with 300 hidden units, and a context of 10 words. 3) The word vectors are extracted from all the selected documents in the Dementia-Bank dataset. 4) Four statistical functionals are computed for the word vectors extracted from each transcript: average, standard deviation, skewness, and kurtosis. Thus, a 1200-dimensional feature vector was formed per transcript.

2.2 Term Frequency—Inverse Document Frequency

The language production impairments exhibited by AD patients also include a high number of non-coherent repetitions and sentences [6]. TF-IDF features represent the relevance of each word in a document, averaged by its global importance in the whole dataset [17]. The objective of TF-IDF features is to model the vocabulary of the patients and the relevance of each word in their transcripts.

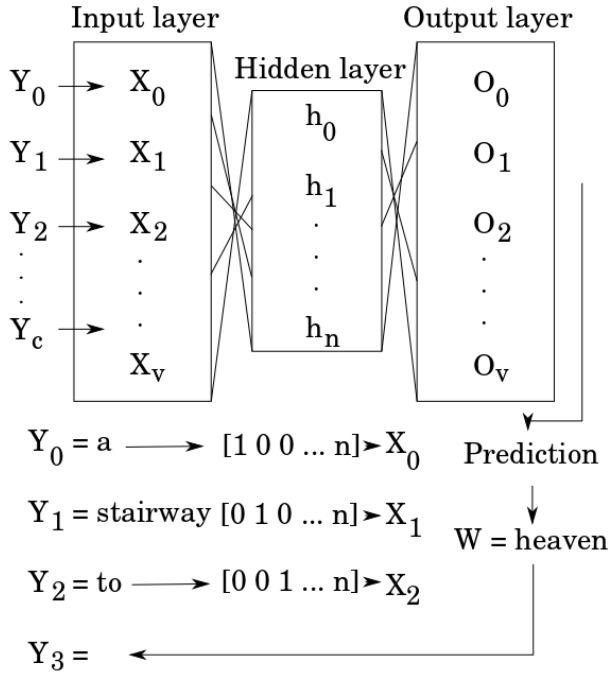

Fig. 2. Description of the CBoW architecture. Source: Created by the authors.

On the one hand, Term Frequency (TF) features of each word in a document are obtained as the ratio between the number of times that the word appears in a document and the total amount of words in said document, according to (1). On the other hand, Inverse Document Frequency (IDF) features of each word are calculated as the logarithm of the total quantity of documents divided by the number of documents that contain that word, according to Expression (2). In (1), $TF_{\omega,t}$ is the TF feature associated with word ω in the transcript t, and $f_{\omega,t}$ is the frequency of ω in t. In (2), IDF_{ω} is the IDF feature of each word ω, T is the total number of transcripts, and T_{ω} is the total number of transcripts where ω is present.

$$TF_{\omega,t} = \frac{f_{\omega,t}}{\sum_t \omega_j} \quad (1)$$

$$IDF_{\omega} = \frac{T}{T_{\omega}} \quad (2)$$

The TF-IDF feature of each word ω is given by (3), and it is the result of computing the product of (1) and (2); this was done for each word ω in the transcripts. A 1408-dimensional feature vector was calculated per transcript; such dimensions were given by the vocabulary size of the Dementia-Bank dataset.

$$TF - IDF_{\omega} = TF_{\omega,t} * IDF_{\omega} \quad (3)$$

2.3 Grammar features

The feature sets studied in this work are inspired by clinical evaluations to assess the neurological state of AD patients. Additionally, we propose grammar features to model the sentence structuring capabilities of AD patients, who show deficits in using nouns and verbs [23]. Moreover, AD patients have problems to use verbs when arguments are involved [24]. The goal of such grammar features is to assess the sentence structuring capabilities of AD patients by counting the elements involved in the structuring of sentences and the number of grammatical elements (such as verbs and nouns) contained in their transcripts.

Eight grammar features were used with their corresponding equations: Readability of the transcript calculated with the Flesch reading score (FR) (4), Flesch-Kincaid grade level (FG) (5), propositional density (PD) (6), and content density (CD) of the transcript (7). The FR score indicates the educational attainment a person needs to easily read a portion of text, ranging from 1 to 100. A score between 70 to 100 means the text is easily readable by a person without specialized education. A score below 30 indicates that a text requires effort and a higher education to be read [21]. The FG measures writing skills, and ranges from 0 to 18. A score of 18 means a very complex and well-structured text. FG scores below 6 indicate a barely elaborated text [21]. PD measures the overall quality of propositions in a text.

In turn, CD quantifies the amount of useful information in a transcript. The constants in (4) and (5) are defined as standard for the English language. The feature set is completed with Part-Of-Speech (POS) counts: Noun to Verb Ratio (NVR), Noun Ratio (NR), Pronoun Ratio (PR) and Subordinated to Coordinated Conjunctions Ratio (SCCR) (8), (9), (10) and (11) [25].

The selected POS counts measure the quality of the syntactical abilities of AD patients when structuring sentences.

$$\text{FR} = 206.835 - 1.015\frac{\#\ words}{\#\ sentences} - 84.6\frac{\#\ syllables}{\#\ words} \quad (4)$$

$$\mathrm{FG} = 0.39\frac{\#\ words}{\#\ sentences} + 11.8\frac{\#\ syllables}{\#\ words} + 15.59 \tag{5}$$

$$PD = \frac{\#(verbs + adjectives + prepositions + conjunctions)}{\#\ words} \tag{6}$$

$$\mathrm{CD} = \frac{\#(verbs + nouns + adjectives + adverbs)}{\#\ words} \tag{7}$$

$$\mathrm{NVR} = \frac{\#\ nouns}{\#\ verbs} \tag{8}$$

$$\mathrm{NR} = \frac{\#\ nouns}{\#\ (nouns + verbs)} \tag{9}$$

$$\mathrm{PR} = \frac{\#\ pronouns}{\#\ (pronouns + nouns)} \tag{10}$$

$$\mathrm{SCCR} = \frac{\#\ (subordinated\ conjunctions)}{\#(coordinated\ conjunctions)} \tag{11}$$

2.4 Data

The clinical Pitt Corpus from the Dementia-Bank dataset was used in this study [10]. The data are the result of a longitudinal study on AD conducted by the University of Pittsburgh School of Medicine. It contains the transcripts of spontaneous speech from HC subjects as well as individuals who possibly and probably have AD. The acquisition of such data involved annual interviews with participants, who described the situations occurring in the Cookie Theft picture (Fig. 3), which is part of the Boston Diagnostic Aphasia Examination.

The participants' verbal utterances in English language were recorded and transcribed.

This study considers data from 98 individuals from the AD group and 98 from the HC group. Participants mentioned their age during the first interview.

Table 1 shows the demographic and clinical information of AD patients and HC. Fig. 4 is a histogram of their Mini-Mental-State-Examination (MMSE) scores with the corresponding probability density distribution of both groups. Fig. 5 presents the box-plot, histogram, and the probability density distribution of the age of both groups.

Table. 1. Demographic and clinical data of patients and controls. Source: Created by the authors.

	AD patients	**HC subjects**
Gender [F/M]	64/34	58/40
Age [F/M]	70.8 (8.4) / 66.5 (7.8)	63.3 (7.9) / 64.6 (7.5)
Educational attainment [F/M]	11.9 (2.3) / 13.4 (2.9)	14.0 (2.5) / 13.8 (2.4)
Years since diagnosis [F/M]	3.6 (1.6) / 3.2 (1.4)	
MMSE [F/M]	20.1 (4.1) / 20.2 (5.2)	29.2 (1.0) / 28.9 (1.1)

Note: The values are expressed as mean (standard deviation). F = female, M = male. Education values are expressed in years

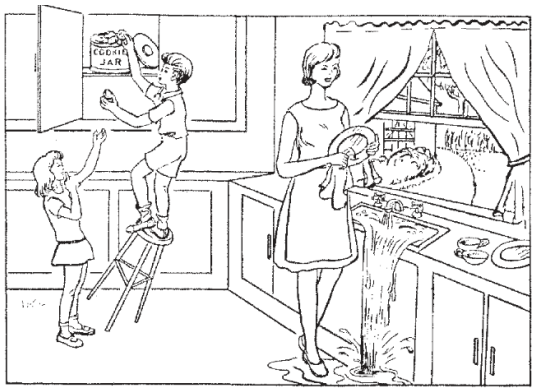

Fig. 3. Cookie Theft picture from the Boston Diagnostic Aphasia Examination. Source: [14].

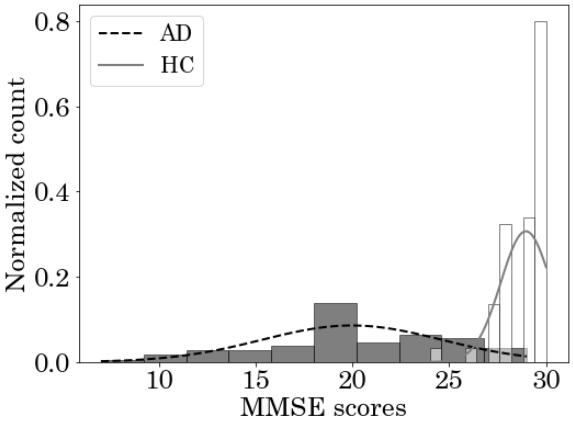

Fig. 4. Histogram of MMSE scores and probability density function of the AD and HC groups
Source: Created by the authors.

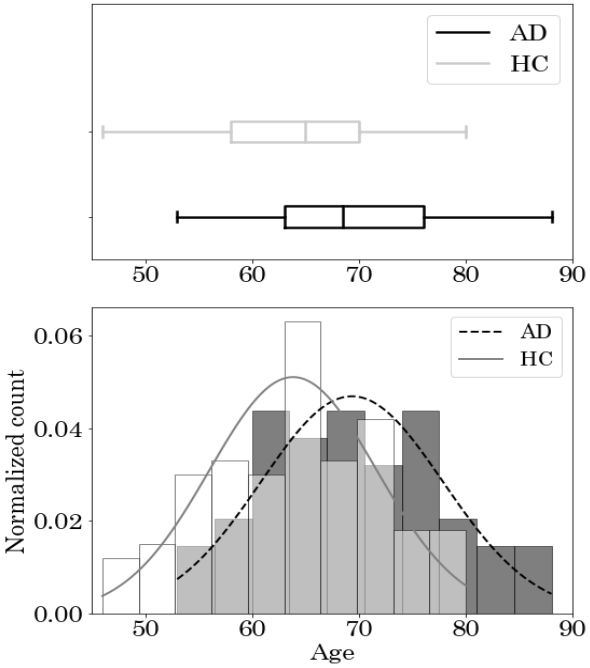

Fig. 5. Box-plot, histogram, and probability density distribution of the ages of the AD and HC groups
Source: Created by the authors.

3. EXPERIMENTS, RESULTS, AND DISCUSSION

3.1 Classification

The feature sets were classified in four experiments using three different methods: SVM, Random Forest (RF), and K-Nearest Neighbors (KNN). Scikit-learn was used to classify the proposed models [26]. Default parameters were used unless specified.

A Leave-One-Out Cross-Validation strategy was followed, and the hyper-parameter optimization was performed via exhaustive grid-search, according to the accuracy obtained in the development set. The objective of such validation strategy is to compare our results with previous studies, such as [14]. The range of the hyper-parameters evaluated in the training process is shown in Table 2. One speaker was used for testing; the rest were divided into 9 groups. Eight groups were used for training, and one group was employed for the hyper-parameter optimization.

3.2 Experiments and Results

Four experiments were conducted. The results of the three classifiers are reported for each experiment. The accuracy (ACC), sensitivity (SEN), specificity (SPEC), and the area under the Receiver Operating Characteristic curve (AUC) of the three classifiers are listed in Table 3.

Experiment 1: W2V features were considered. The best classifier in this experiment was SVM with a linear kernel (ACC = 81.3 %). The results of the Mcnemar test show a significant difference between the best result and those obtained with other classifiers ($p<<0.02$).

Experiment 2: TF-IDF features were considered. In this case, the best classifier was the SVM with linear kernel (ACC = 81.6 %), and the results of the Mcnemar test also show a significant difference between the best result and those obtained with the other classifiers ($p<<0.001$).

Experiment 3: Grammar features were considered. The best classifier for this feature set was RF (ACC = 72.8 %). The results of the Mcnemar test show a significant difference between the best result and those obtained with the other classifiers ($p<<0.01$).

Experiment 4: The early fusion of the feature sets was considered. In this case, the best classifier was RF (ACC = 81.7 %). Once more, there is a significant difference between the best results and those obtained with the other classifiers (p-value approx. 10^{-12}).

A statistical comparison between the best results obtained with the early fusion strategy and those obtained with each feature set reveals significant differences. Table 4 shows the p-values obtained in the experiments.

Table. 2. Range of the hyper-parameters used to train the classifiers. Source: Created by the authors.

Classifier	Parameter	Values
SVM	Kernel	{Linear, RBF}
	C	$\{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$
	γ	$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$
KNN	n_neighbors	{3,5,7,9,11,15}
RF	max_depth	{1,3,5,7}
	n_estimators	{5,20,30,50,100}
	min_samples_leaf	{1,2,4,10}
	bootstrap	{True, False}

Table. 3. Results obtained with different feature sets and classifiers to identify AD patients and HC subjects
Source: Created by the authors.

Experiment	Features	Classifier	ACC (%)	SEN (%)	SPEC (%)	AUC	Best parameters
1	W2V	SVM (linear)	81.3 ± 3.6	77.0	83.3	0.80	$C: 10^{-4}$
		SVM (RBF)	80.5 ± 3.4	79.1	81.6	0.79	C: 1 γ: 10^{-4}
		KNN	77.9 ± 3.5	75.0	75.1	0.75	$n\_neighbors$: 13
		RF	79.7 ± 3.4	77.0	81.1	0.79	max_depth: 7 $min\_samples\_leaf$: 4 $n\_estimators$: 100 $bootstrap$: $False$
2	TF-IDF	SVM (linear)	81.6 ± 3.6	78.0	79.3	0.79	C: 10^{-3}
		SVM (RBF)	80.9 ± 3.3	77.1	78.3	0.78	C: 10 γ: 10^{-4}
		KNN	76.7 ± 3.6	75.0	76.9	0.76	$n\_neighbors$: 13
		RF	81.1 ± 3.3	78.1	79.5	0.79	max_depth: 7 $min\_samples\_leaf$: 4 $n\_estimators$: 100 $bootstrap$: $False$
3	Grammar	SVM (linear)	70.6 ± 8.8	68.9	65.9	0.66	$C: 10^{-2}$
		SVM (RBF)	72.5 ± 1.0	72.2	72.0	0.70	C: 1 γ: 10^{-2}
		KNN	70.8 ± 8.1	72.0	67.0	0.68	$n\_neighbors$: 7
		RF	72.8 ± 3.7	73.0	73.5	0.71	max_depth: 3 $min\_samples\_leaf$: 10 $n\_estimators$: 100 $bootstrap$: $False$
4	**Fusion**	SVM (linear)	81.3 ± 3.6	77.0	84.0	0.80	C: 10^{-4}
		SVM (RBF)	80.5 ± 3.4	77.0	81.0	0.79	C: 10 γ: 10^{-4}
		KNN	77.0 ± 3.5	74.7	73.1	0.74	$n\_neighbors$: 13
		RF	$\mathbf{81.7 \pm 3.4}$	**78.4**	**81.7**	**0.80**	$\boldsymbol{max\_depth}$: **7** $\boldsymbol{min\_samples\_leaf}$: **4** $\boldsymbol{n\_estimators}$: **100** $\boldsymbol{bootstrap}$: $\boldsymbol{False}$

Table. 4. Comparison of p-values (obtained with the Mcnemar test) between the best results of experiments 1–3 and experiment 4
Source: Created by the authors.

Experiment			**p-value**
Early fusion	vs.	TF-IDF	$\approx 1.73x10^{-7}$
		Grammar	$\approx 1.30x10^{-14}$
		W2V	≈ 0.005

3.3 Discussion

According to the results, the model based on the combination of the three feature sets and the RF classifier is the most accurate to classify AD patients and HC subjects. The values obtained with the linear SVM indicate that most of the extracted features are linearly separable. The accuracy obtained with this classifier ranges from 78.3 % to 85.1 %. The TF-IDF and W2V models exhibited similar results in general. According to the high and balanced values of specificity and sensitivity, and in spite of the high values of the MMSE scores of several AD patients, the proposed approach seems to be accurate and robust. Additionally, the grammar features are highly accurate (72.8 %) and effective. The reduced number of features in the grammar set indicates that this approach is suitable and promising.

It is important to highlight the results of the statistical information of all the experiments. There is a weak statistical relationship between the predictions of all the classifiers in the experiments, which means that the errors and correct predictions of each classifier were different. In experiment 4, the classifiers exhibit the weakest statistical relationship as a result of the early fusion of the feature sets. The accuracy values obtained using the early fusion strategy show improvements in the RF classifier, which is the most benefited with the combination of feature sets. SVM and KNN classifiers showed no significant improvement in performance compared with experiments 1 and 2.

The results of this study can be directly compared to those in [9] and [14], since we adopted the same cross-validation strategy. These results, are slightly lower than those in related studies, however, those reported there could be optimistic, since the features extracted in the BOW model and n-gram models were computed with information of the vocabulary of the test set. In a more realistic clinical environment, a more general feature set, such as W2V, is preferred. TF-IDF features showed an important role in modeling the difficulties of AD patients to find appropriate words. In addition, grammar features proved to be an alternative to detect, without a complex feature extraction process, AD patients' impairments to structure sentences with useful information.

4. CONCLUSIONS

This study used to word-embed features (i.e., statistical functional, TF-IDF features, and grammar features) to classify AD patients and HC subjects in the Pitt Corpus of the Dementia-Bank dataset employing different classifiers. A total of 1200 word-embedding features, 1408 TF-IDF features, and 8 grammar features were computed based on the transcripts in the dataset. Each feature set was classified separately. An early fusion strategy of the three feature sets was also considered.

The language impairments of AD patients were successfully modeled using the proposed methods. TF-IDF features modeled the deteriorating vocabulary and low word relevance in the transcripts of AD patients. Semantic, comprehension, and memory loss impairments of AD patients were modeled with W2V features.

The sentence structuring capabilities of AD patients were modeled with grammar features. All the models achieved high accuracies in the automatic discrimination between AD patients and HC subjects. The models obtained from the W2V feature set and the TF-IDF feature set showed a similar performance, although the early fusion strategy contributed to a better model. The early fusion achieved accuracies of up to 81.7 %. When only grammar features were considered, the proposed approach exhibited accuracies of

up to 72.8 %. The features based on the W2V model showed a significant importance as the embeddings were extracted from a non-specialized knowledge database rather than the classification dataset. TF-IDF features were extracted directly from the transcripts used in this work, and they have a higher dimensionality than the W2V model. Grammar features were found to be important and produced promising results without the need of complex calculations in the extraction process.

We believe that further experiments can be designed to identify the most suitable features for clinical evaluations.

Statistical differences between the classifiers were found in all the experiments. This suggests that the experiments that use an ensemble or stacking techniques could produce better results. Experiments with deep learning techniques, a bigger dataset to retrieve TF-IDF features, a larger word vector dimension, and a considerably larger set of grammar features are needed in future work. Additionally, word embeddings of novel language models based on more sophisticated neural network architectures, such as BERT and XLNET, could lead to better results because such models have achieved state-of-the-art performance in numerous NLP tasks [27], [28], [29]. Further research is required with the aim of finding possible clinical interpretations to the results based on these kinds of models.

5. ACKNOWLEDGMENT

This work was funded by Comité para el Desarrollo de la Investigación (CODI) from Universidad de Antioquia grant No. 2017-15530 and EU Marie Curie programme grant No. 766287.

6. REFERENCES

[1] S. R. Chandra, "Alzheimer's disease: An alternative approach", *Indian J. Med. Res.*, vol. 145, no. 6, pp. 723 - 729, Jun. 2017. https://doi.org/10.4103/ijmr.IJMR_74_17

[2] C. M. Henstridge, B. T. Hyman, and T. L. Spires-Jones, "Beyond the neuron–cellular interactions early in Alzheimer disease pathogenesis", *Nature Reviews Neuroscience*, vol. 20, pp. 94-108, Jan. 2019. https://doi.org/10.1038/s41583-018-0113-1

[3] F. J. Huff, J. T. Becker, S. H. Belle, R. D. Nebes, A. L. Holland, and F. Boller, "Cognitive deficits and clinical diagnosis of Alzheimer's disease," *Neurology*, vol. 37, no. 7, pp. 1119–1124, Jul. 1987. https://doi.org/10.1212/WNL.37.7.1119

[4] J. A. Small, S. Kemper, and K. Lyons, "Sentence comprehension in Alzheimer's disease: Effects of grammatical complexity, speech rate, and repetition," *Psychol and Aging*, vol. 12, no. 1, pp. 3–11, Mar. 1997. https://doi.org/10.1037/0882-7974.12.1.3

[5] M. Nicholas, L. K. Obler, M. L. Albert, and N. Helm-Estabrooks, "Empty Speech in Alzheimer's Disease and Fluent Aphasia," *J. Speech, Lang. Hear. Res.*, vol. 28, no. 3, pp. 405 - 410, Sep. 1985. https://doi.org/10.1044/jshr.2803.405

[6] B. E. Murdoch, H. J. Chenery, V. Wilks, and R. S. Boyle, "Language disorders in dementia of the Alzheimer type," *Brain and Language.*, vol. 31, no. 1, pp. 122 - 137, May. 1987. https://doi.org/10.1016/0093-934X(87)90064-2

[7] D. A. Snowdon, S. J. Kemper, J. A. Mortimer, L. H. Greiner, D. R. Wekstein, and W. R. Markesbery., "Linguistic Ability in Early Life and Cognitive Function and Alzheimer's Disease in Late Life: Findings From the Nun Study", *JAMA clinical challenge*, vol. 275, no. 7, pp. 528 - 532, Feb. 1996. https://doi.org/10.1001/jama.1996.03530310034029

[8] A. Almor, D. Kempler, M. C. MacDonald, E. S. Andersen, and L. K. Tyler, "Why Do Alzheimer Patients Have Difficulty with Pronouns? Working Memory, Semantics, and Reference in Comprehension and Production in Alzheimer's Disease", *Brain and Language.*, vol. 67, no. 3, pp. 202 - 227, May. 1999. https://doi.org/10.1006/brln.1999.2055

[9] S. O. Orimaye, J. S.-M. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri, "Predicting probable Alzheimer's disease using linguistic deficits and biomarkers", *BMC Bioinformatics*, vol. 18, no. 34, Jan. 2017. https://doi.org/10.1186/s12859-016-1456-0

[10] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The Natural History of Alzheimer's Disease: Description of Study Cohort and Accuracy of Diagnosis", *Arch. Neurol.*, vol. 51, no. 6, pp. 585 - 594, Jun. 1994. https://doi.org/10.1001/archneur.1994.00540180063015

[11] P. F. Brown, P. V DeSouza, R. L. Mercer, V. J. Della Pietra, and J. C. Lai, "Class-based *n*-gram Models of Natural Language," *Computational. Linguists.*, vol. 18, no. 4, pp. 467–479, Dec. 1992. Available: URL

[12] B. Mirheidari, D. Blackburn, T. Walker, A. Venneri, M. Reuber, and H. Christensen, "Detecting Signs of Dementia Using Word Vector Representations," in *Interspeech,* Hyderabad, 2018, pp. 1893 -1897. https://doi.org/10.21437/Interspeech.2018-1764

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in *26th International Conference on Neural Information Processing Systems*, Nevada, 2013, pp. 3111 - 3119. Available: URL

[14] P. Klumpp, J. Fritsch, and E. Nöth, "ANN-based Alzheimer's disease classification from bag of words", in *Speech Communication; 13th ITG-Symposium*, Oldenburg, 2018. pp. 1-4. Available: URL

[15] A. Budhkar and F. Rudzicz, "Augmenting word2vec with latent Dirichlet allocation within a clinical application", in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, 2019, pp. 4095-4099. https://doi.org/10.18653/v1/N19-1414

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space", in *Proceedings of the International Conference on Learning Representations*, Arizona, 2013, pp.1-12. Available: URL

[17] G. Salton, and M. J. McGill, *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1986. Available: URL

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, vol. 3, pp. 993 - 1022, Jan. 2003. Availabe in: URL

[19] P. Herd, D. Carr, and C. Roan, "Cohort Profile: Wisconsin longitudinal study (WLS)," *International journal of epidemiology*, vol. 43, no. 1, pp. 34 - 41, Feb. 2014. https://doi.org/10.1093/ije/dys194

[20] A. Pistono, M. Jucla, C. Bézy, B. Lemesle, J. Le Men, and J. Pariente, "Discourse macrolinguistic impairment as a marker of linguistic and extralinguistic functions decline in early Alzheimer's disease," *Int. J. Lang. Commun. Disord.*, vol. 54, no. 3, pp. 390 - 400, May. 2019. https://doi.org/10.1111/1460-6984.12444

[21] R. Rehrek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora" in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, 2010, pp. 45–50. Available: URL

[22] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, United States: O'Reilly Media, Inc. 2009. Available: URL

[23] A. Almor *et al.*, "A common mechanism in verb and noun naming deficits in Alzheimer's patients," *Brain and Language*, vol. 111, no. 1, pp. 8 -19, Oct. 2009. https://doi.org/10.1016/j.bandl.2009.07.009

[24] M. Kim and C. K. Thompson, "Verb deficits in Alzheimer's disease and agrammatism: Implications for lexical organization," *Brain and Languaje*, vol. 88, no. 1, pp. 1-20, Jan. 2004. https://doi.org/10.1016/S0093-934X(03)00147-0

[25] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," Naval Technical Training Command Millington TN Research Branch Report, United States, IST technical report, 1975. Available: URL

[26] C. Roth, "Boston Diagnostic Aphasia Examination", in *Encyclopedia of Clinical Neuropsychology*, 3st ed, New York: Springer New York, 2011. pp. 338 - 468 https://doi.org/10.1007/978-0-387-79948-3_868

[27] F. Pedregosa, et al., "Scikit-learn: Machine Learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830. Oct. 2011. Available: URL

[28] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: "Pre-training of Deep Bidirectional Transformers for Language Understanding", in *Conference of the North American Chapter of the Association for Computational Linguistics,* Minneapolis, 2019, pp. 1-16. Available: URL

[29] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in *Conference of the North American Chapter of the Association for Computational Linguistics,* Minneapolis, 2019, Available: URL