IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 July 2022, accepted 17 August 2022, date of publication 31 August 2022, date of current version 8 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3203068

RESEARCH ARTICLE

An Evaluation on Information Composition in Dementia Detection Based on Speech

CHUHENG ZHENG[1], (Member, IEEE), MONDHER BOUAZIZI[2], (Member, IEEE), AND TOMOAKI OHTSUKI[2], (Senior Member, IEEE)

[1]Graduate School of Science and Technology, Keio University, Yokohama, Kanagawa 223-8522, Japan
[2]Faculty of Science and Technology, Keio University, Yokohama, Kanagawa 223-8522, Japan

Corresponding author: Tomoaki Ohtsuki (ohtsuki@keio.jp)

This research was conducted and the herein results were achieved with support by a grant funded by Eisai Co., Ltd., JAPAN.

ABSTRACT In recent years, scientists are paying much attention to the research on automatic dementia detection that could be applied to the speech samples of dementia patients. In a related context, recent research has seen the fast development of Deep Learning (DL) and Natural Language Processing (NLP). The techniques developed for text classification or sentiment analysis have been applied to the field of early dementia detection by many researchers. However, text classification and sentiment analysis are different tasks from dementia detection, which makes us believe that for dementia detection, some adjustments would help improve the performance of the machine learning models. In this work, we implemented experiments with various language models including traditional *n*-gram language models, Average stochastic gradient descent Weight-Dropped Long Short-Term Memory (AWD-LSTM) models, and attention-based models to evaluate the speech data of dementia patients. Unlike traditional works where the text is stripped from stop words, we propose the idea of exploiting the stop words themselves, since they offer non-context information which helps to identify dementia. As a result, 3 different language models are prepared in this work: a model processing only context words, a model processing stop words and Part-of-Speech (PoS) tag sequences, and a model processing both of them. By performing the aforementioned experiments, we show that both grammar and vocabulary contribute equally to classification: The 3 models achieve an accuracy equal to 70.00%, 76.16%, and 81.54%, respectively.

INDEX TERMS Dementia detection, deep learning, language models, transfer learning, natural language processing.

I. INTRODUCTION

A. BACKGROUND

Dementia belongs to the category of neural degenerative diseases that cause deterioration of cognitive functioning gradually in a long term. Dementia usually has a severe influence on language ability, memory, and executive functions. It also leads to a lack of motivation, motor problems, and emotional distress. With the development of the disease, these symptoms become increasingly severe, which reduces the autonomy of the patients as well as their well-being and that of their caregivers [1].

With the age being the main risk for Alzheimer's disease which accounts for the majority of dementia patients, the number of dementia patients is expected to increase in the following years because the population over 65 years old is predicted to triple between 2000 and 2050 [2]. As such, dementia is expected to have an ever-growing immense impact on society. In 2015, the estimated number of dementia patients worldwide is over 47.5 million. According to World Health Organization (WHO) [3], a longitudinal study where the researchers keep tracking the status of the subjects through the years finds the annual incidence of dementia is between 10 and 15 cases per thousand people. Patients who developed dementia on average have 7 years of life expectancy and less than 3% of dementia patients would live longer than 14 years or more [3].

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

This severe situation is calling the institutions and researchers to put more effort on dementia prevention and early detection. Cost-effective and scalable methods for detection of dementia that can capture the subtle symptoms from the pre-clinical stages, such as subjective memory loss, or worse conditions like Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD).

Detection and prevention as early as possible are proven effective to improve the therapy effectiveness and quality of life of the patients. The neuropathology of AD consists of several phenomena: intracellular accumulation of tau-protein fibers and extracellular accumulation of beta-amyloid plaques. These symptoms are noticed to start silently up to 20 years before a subject is observed to show obvious symptoms, at which stage, treatment of the patients becomes futile.

Memory, attention, language, and decision-making are components of cognition. MCI could be a sign of potentially developing Alzheimer's disease, while severe cognitive impairment could be a sign of the presence of dementia. It is now widely recognized that cognition plays an important role in sustaining the autonomy of seniors. At the same time, neurodegenerative diseases, particularly AD, that would cause cognitive impairment has become a great concern in public health care.

Developing an effective automatic speech analysis system is more psychologically accepted when the subject is aided by a real person or an interactive robot. Ambient sensors or devices that cannot provide meaningful interaction are less acceptable [4]. Hence, this kind of system usually involves research in natural communication instead of fixed text [5].

From a medical perspective, AD would disrupt patients' ability to follow conversations, even the simplest daily instructions. However, this symptom is not obvious in scripted talk, which makes the information richness in the scripted talk less dense than that of a spontaneous conversation [6].

Research about early dementia detection has received intensive attention in recent years. This is because to establish effective prevention measures for AD, it is necessary to detect AD pathology several years before the patient shows clinical symptoms [7]. Currently, image-based techniques like Positron Emission Tomography (PET) scan or Magnetic Resonance Imaging (MRI) scan and cerebrospinal fluid analysis provide an accurate diagnosis.

Among different types of data, speech is considered a valuable source of clinical information. Human speech has a close relation to cognitive status and is used as the basic information source for a lot of applications used for mental health assessment. The language patterns are related to the cognitive status and reflect the decline of cognitive functioning. Thus, it could be used in the design of assistive technologies [8]. For one thing, dementia usually causes language impairment, which is shown by difficulties in word-finding, understanding, accuracy, and lack of coherence in speech [1]. Furthermore, language also relies on other cognitive functions including executive functions so that communication happens in a sound and meaningful way. Cognitive functions also play important roles in decision making, strategy planning, and problem-solving, which are significant to communications [9]. Speech data are also common and easy to collect. In the past few years, using Natural Language Processing (NLP) and machine learning techniques to detect dementia based on speech and language data is receiving attention from researchers around the world [9].

Language is a good indicator for early dementia detection. However, analyzing the language is difficult, challenging, and time-consuming because it requires the involvement of manual analysis performed by professionals. The advances in speech and language analysis techniques are bringing us 3-fold advantages. First, it could help to develop reliable tools for detecting the differences between dementia speech samples and non-dementia speech samples. Besides, it can quantify the stages of dementia. It also can distinguish between different types of dementia [9], [10], [11].

From a medical perspective, dementia is not a single disease. The term applies to a wide spectrum of medical disorders. AD accounts for more than 60% of dementia cases [12], [13]. Even while certain dementia disorders may be healed if discovered early enough, the vast majority of dementia diseases are incurable. Expert evaluation and early diagnosis of dementia symptoms, however, may help to halt the advancement of the disease. Another merit of the early detection of dementia is that it largely helps others around the patient better understand the patient's previously puzzling behavior. Scientists are paying increasing attention to dementia diagnosis and developing novel ways for identifying it due to its importance. As a consequence, various research works in the past had focused on dementia detection [14], [15], [16]. Dementia testing may take several forms, ranging from cognitive and brain imaging to laboratory testing and brain scans [17], [18]. However, these techniques are usually expensive and time-consuming to implement. This work might be automated to save money and make it more accessible to the general population.

As a result, the scientific community has been looking at numerous ways to execute the work of dementia diagnosis automatically. Automation of dementia diagnosis utilizing cutting-edge Artificial Intelligence (AI) technologies, in particular, might make this activity considerably more economical and accessible. This is because AI technologies have made a few advances in recent years, allowing it to recognize small patterns in a range of data formats while also being substantially less expensive [19], [20], [21].

Concerning the topic of dementia detection, a few data sets have been publicly available to experiment with, such as DementiaBank[1] and Dem@Care.[2] They present data in different formats, notably audio, video, and transcribed text of dementia patients and control subjects. Among these, the

[1]https://dementia.talkbank.org/
[2]https://demcare.eu/datasets/

speech format, whether in audio or text format, is a very informative type of information that has attracted the most attention. Many works have addressed the idea of processing the text in its transcribed format or as an audio signal for dementia detection [11], [22], [23], thanks to the advances in the field of NLP as well as audio processing. For instance, a wide variety of techniques related to text classification have been proposed in the literature [11], [24]. Whether the task is sentiment analysis, hate speech detection, or automated bots identification [25], the overall way to perform the task is roughly the same: extract clues from the text itself and use Artificial Intelligence (AI), namely machine learning and deep learning to identify the target class. Applying these techniques in dementia detection has led to some promising results [26].

However, we believe that the distinctions between opinion mining and dementia diagnosis are significant when using text classification approaches. The substance of the text and the meaning of the words include the majority of the information required to conduct tasks like sentiment analysis and hate speech identification. Yet, this is not always the case when it comes to dementia detection.

B. RELATED WORK

Due to the factors that cause dementia and the symptoms shown in the patients are multimodal, the past years have seen different approaches being proposed for early dementia detection.

A study by Roark [27] annotated a few speech features and aligned them by using NLP and automatic speech recognition tools. The same features are also extracted manually by human annotation. They studied the differences in the performance between the automated feature extraction and its human-annotated version. They evaluated 74 speech recordings to classify between Mild Cognitive Impairment (MCI) and healthy subjects. Their model of the best performance reached an Area Under the Curve (AUC) of 0.86 by combining speech and language features and cognitive test scores.

Zhu *et al.* [6] explored different transfer learning techniques for dementia detection, which involved using pre-trained models and fine-tuning them on the dementia data set.

Jarrold *et al.* [28] used a data set that included semi-structured interviews from 9 healthy individuals, 9 individuals with Alzheimer's disease, 9 individuals with frontotemporal dementia, 13 individuals with semantic dementia, and 8 individuals with progressive nonfluent aphasia. They retrieved 41 features using an Automatic Speech Recognition (ASR) system, including speech rate, the mean and standard deviation of pause, vowel, and consonant length. Using a multilayered perceptron network, they were able to achieve an 88% classification accuracy for AD vs. healthy participants based on lexical and auditory data.

Luz *et al.*, in their recent research [23], extracted features based on a graph that encodes turn pattern and speech rate from the Carolina Conversations Collection [29], which includes recordings of interviews with dementia patients and healthy people. By using these features, they composed an additive logistic regression model that could distinguish speech between healthy subjects and dementia patients.

In some studies, signal processing, and NLP techniques are used to detect signs of dementia that may be imperceptible to human professionals. For example, Tóth *et al.* [30] discovered that even though human annotators could not recognize pauses (sounds like "hmmm," etc.) reliably, these features are easy to collect with an ASR system. In this research, several acoustic parameters (hesitation ratio, speech speed, length, number of silent and filled pauses, and duration of utterance) were extracted from the recorded speech of 38 healthy controls subjects and 48 patients with MCI talking about two short films. They found that ASR-extracted features outperformed manually computed features (69.1% accuracy) when combined with machine learning approaches, notably with a Random Forest classifier (75% accuracy). König *et al.* [31] employed a similar machine learning approaches and showed an accuracy of 79% when discriminating MCI individuals from healthy counterparts, 94% for AD vs. healthy, and 80% for MCI vs. AD. Their tests, on the other hand, were conducted on non-spontaneous speech data collected under controlled settings as part of a neuropsychological evaluation that also included mechanically transcribed text.

The idea of combining two perplexity values, one from a language model trained on speech samples of dementia, and one from a language model trained on speech samples of healthy, was proposed by Wankerl *et al.* [32]. Perplexity is used to estimate the fit between a probabilistic language model, and a sample of previously unseen text in the training.

The n-gram language model is a method widely used in processing speech or written language [33]. N-gram language models create probability density from training text data by calculating the frequencies of the word sequences. In the simplest uni-gram/1-gram language model, the sequence only contains one word. The model counts the words in the training data and assigns a probability to them. For a sentence S:

$$S = (w_1, w_2, \ldots, w_k),$$

$w_1, w_2, \ldots, w_k$ represent the 1st word, 2nd word, …, k-th word in the sentence S. For any sentence in the test data, the model estimates its possibility of existence based on the training data. In the case of the uni-gram language model, the sentence probability $p(S)$ equals the product of each word's probability $\prod_{i=1}^{k} p(w_i)$.

The uni-gram language model cannot comprise any contextual information because it only gives the probability distribution of individual words. On the other hand, calculating the probability distribution of individual sentences leads to a unique probability. It might be hard for the model to make proper predictions for new unseen data. Therefore, the length of sequences is limited to a certain small number. A model that calculates the probability distribution of sequences composed of n words is called the n-gram language model. For

example, the tri-gram language model calculates the probability distribution of sequences composed of 3 words. For a sentence $S = (w_1, w_2, \ldots, w_k)$ of k words, when the sequence length is n, the probability is evaluated by [32]:

$$p(S) = \prod_{i=1}^{k} p(w_i | w_{i-n+1}, \ldots, w_{i-1}). \quad (1)$$

In training the n-gram language model, each sentence is padded with special tokens to indicate the beginning and end of the sentence. The token also helps to calculate the probability of the first word in the sentence. In the case where the data is limited, many sequences in the test data may not appear in the training data. Therefore, it is important to introduce a smoothing method into the n-gram language models. Additive smoothing [34] simply assigns a constant probability for the sequence that did not appear in the training data.

As the first step in the work [32], two language models are created from all the data from dementia subjects (LM_{dem}) and all the data from the healthy subjects (LM_{con}). These two models represent the typical speech patterns of dementia speech and non-dementia speech. To evaluate the speech of each participant in the data set, an additional model needs to be created so that the test data itself does not appear in the training data set. For subjects with more than one recording sample, their speech samples are entirely excluded from the data set. This is necessary because each individual might use similar verbalization or re-occurring phrases that are not common or universal to the entire data set which might distort the perplexity distribution.

In the work [32], a cross-validation approach is applied. For each subject, s in the dementia category, a tri-gram model LM_{-s} is created that takes all the speech samples as the training data except those that belong to the subject s. The perplexity p_{own} of every speech from s is calculated using LM_{-s}. While p_{other} is obtained using the model LM_{con}, which certainly does not contain any recording from subject s because they are in different categories. The cross-validation is repeated for all the subjects in the data set. In addition to p_{other}, p_{own}, the difference between them is added as another feature, which is calculated in the following:

$$p_{\text{diff}} = \begin{cases} p_{\text{own}} - p_{\text{other}}, & \text{if } s \in \text{AD Group} . \\ p_{\text{other}} - p_{\text{own}}, & \text{if } s \in \text{Control Group} . \end{cases} \quad (2)$$

By setting a threshold $p_{\text{diff}} = -1.41$, under equal error rate of both categories, they calculated the classification accuracy which equals 77.1%.

Fritsch *et al.* [35] improved the two perplexities methods by introducing neural networks to replace the n-gram language models in the original methods. Instead of conventional statistical language models, they trained Long Short-Term Memory (LSTM) neural network-based language models. They used the two perplexities methods with LSTM and achieved an accuracy of 85.6%.

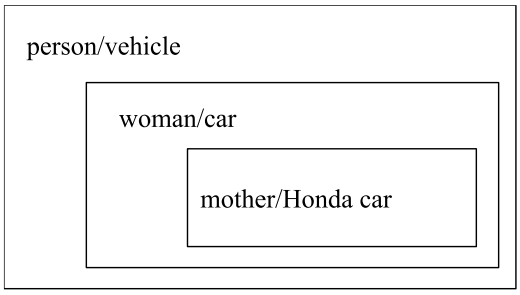

FIGURE 1. Examples of words with different lexical frequencies.

Cohen *et al.* [36] interrogated the two perplexities methods by using artificially synthesized speech data that are created to simulate progressive dementia detection. Bird *et al.* [37] created synthetic narratives by creating a baseline sample and removing and/or replacing the nouns and verbs with higher lexical frequency (mother vs. woman vs. person). Lexical frequency shows how specific a word is in describing the context information. In Fig. 1, we give two groups of examples. In both examples, the words in the outer circles have broader meanings and that includes the meanings of the words shown in the inner circles. Cohen *et al.* [36] followed the work of Bird *et al.* [37] and implemented the two perplexities methods by comparing the original data (words are not replaced nor removed) and the modified data (some words are replaced with higher lexical frequency words). By doing so, they noticed that the perplexity distribution is highly influenced by words' lexical frequency. Their research confirmed that the lexical frequency of vocabulary is effective in detecting dementia.

Previous works analyzed language models and data. However, the analysis does not answer all questions about the topic. They found that language models' perplexities are associated with lexical frequency, but is it the primary information in the detection? Which one contributes the most to the neural network classifiers, syntax, or semantic aspects of the language? What kind of information composition or format do the neural networks take to improve the accuracy performance in dementia detection? When it comes to the medicine area, data is often limited and related to the personal privacy of the patients. Therefore, not all the data is available in the desired amount and form. This results in that deep learning often cannot reach its best performance. Hence, answering these questions helps us develop more reliable, explainable, and accurate models by manually manipulating the data we have. Besides, using language as a source for dementia diagnosis manually is common in traditional methods. This research also aims to provide a new viewpoint for the manual analysis methods, like which part or which component of the sentence deserves more attention.

In this work, we first explored whether the richness, specificity, or variety of the vocabulary along with the difficulty to predict the next word should be the primary indicators in the task of dementia detection. Or if the text's grammatical structure may be a better indicator of dementia detection. We re-implemented the two perplexities methods with

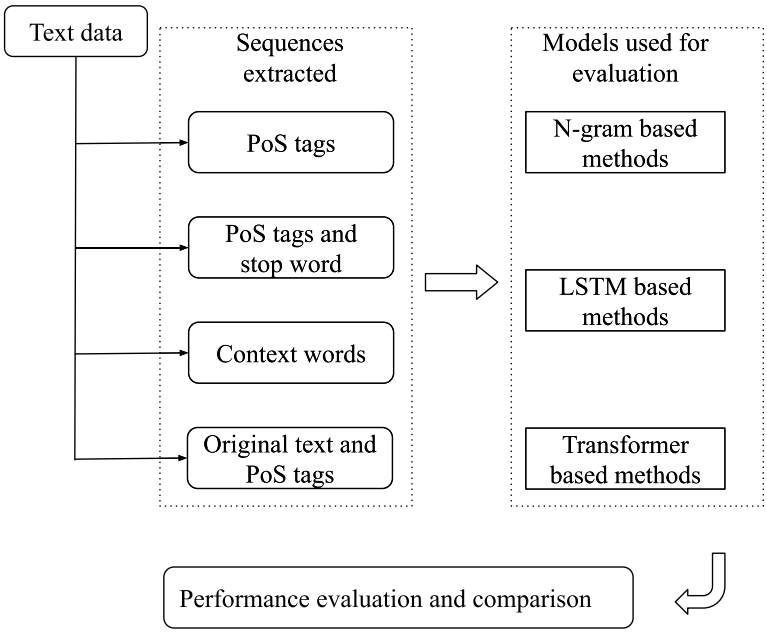

FIGURE 2. Overall pipeline of the research.

Part-of-Speech (PoS) tags and stop word sequences. A PoS tag is a category to which a word is assigned according to its syntactic functions. For example, some simplified and regular PoS tags in English are verb, noun, adjective, etc. Stop words are words that do not contain contextual information, nor do they indicate the context or meaning of the sentences.

Using PoS tags reduces the computation complexity. It also helps to improve the generality of the n-gram language models. In the original methods, they used every word in the data set, but in our implementation, only 33 PoS tags and 127 stop words are used. This can largely decrease the possibility of unknown sequences in the test data and increases the generality of the models. It enables the n-gram models to utilize the syntax of texts in an explicit way, especially in the case where the data set is limited. We also explored using fewer complex models to perform the classification task. Less complex models are easier to train and less likely to have overfitting problems.

Besides, we created multiple classifier models based on different classifier architectures and evaluated the performance using different information compositions. We separate syntax and semantic components from the sentences by incorporating the PoS tags and stop words in the pre-processing of the data set. We researched what's the most efficient information input for the machine learning classifiers. We implemented two main neural network architectures: an LSTM-based neural network classifier built from scratch, and another pre-trained language model-based classifier to do the classification. For both architectures, we train 3 different models:

- a model trained with only context words.
- a model trained with sequences composed of PoS tags of the words, which contain the sentence patterns information but no context information.
- a model trained with sequences composed of both stop words and PoS tags, which contain the sentence patterns information but with finer details.

Lastly, we discussed whether vocabulary variety or richness should be the primary indicator for dementia detection. The overall pipeline of the research is shown in Fig. 2. In the work [36], they evaluated why the two perplexities methods work. They found that perplexities of neural network models are associated with lexical frequencies. We researched more about this topic by the aforementioned methods and showed by experiments that despite the importance of the context words, they are not necessarily the most valuable indicators for dementia detection.

The major contributions of our work are summarized as follows:

- We re-implemented the two perplexities methods with PoS tags and stop word sequences, which costs less computation and has better generality.
- We created multiple classifier models based on different classifier architectures and different information composition to perform dementia detection with an accuracy of 81.54%.
- We discussed whether vocabulary variety or richness should be the primary indicator for dementia detection and showed by experiments that sentence patterns and grammatical fluency are equally important indicators for dementia detection.

II. PROPOSED METHODS

A. DATA SET

The DementiaBank data set provided in TalkBank [38] is used in this work to evaluate the performance of the different introduced models in dementia detection. TalkBank is a

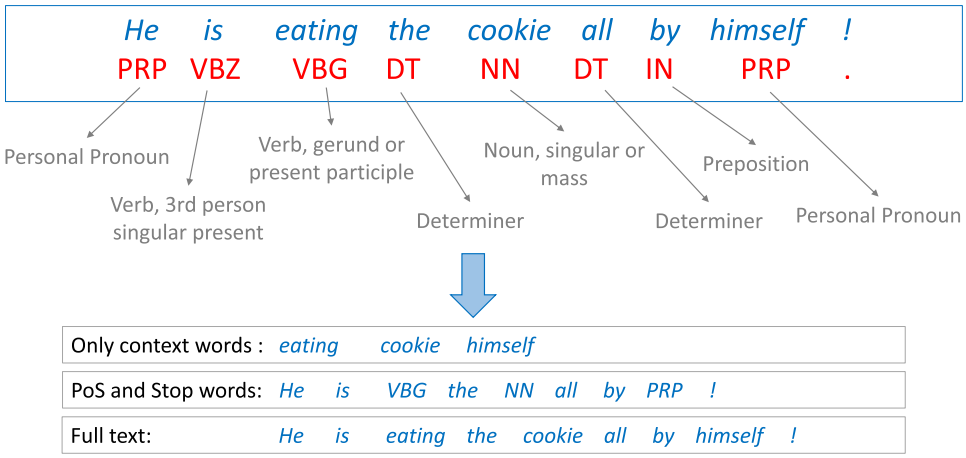

FIGURE 3. Flow Chart.

multi-lingual data set established in 2002 whose object is to encourage research in human and animal communication. It is composed of data in several fields like first/second language acquisition, conversation analysis, dementia, etc. As one of its sub-sets, the DementiaBank is a set of video and audio data with their respective transcribed texts in different languages from dementia patients and non-dementia subjects. Specifically, we use the Pitt Corpus, which includes recorded audio samples with their manually annotated transcripts in a picture description task. The patients were requested to see a picture and describe the content of the picture with the help of an interviewer. It is believed that spontaneous speech is rich in information that could indicate the mental status and cognitive functions of the subjects. With the Pitt Corpus data set, we could detect dementia by evaluating their speech data with machine learning techniques. In this work, we limited our use of the data set to the Pitt Corpus, which contains 309 recordings from 166 Alzheimer's Disease (AD) patients and 242 recordings from 94 healthy subjects. The average number of words is 91 in the AD patients' text samples and 97 in the healthy subjects' text samples.

B. PRE-PROCESSING OF DATA SET

Instead of using the original data set, we extracted 4 kinds of information from the original text. The text in the Pitt Corpus is processed with the natural language parsing tools as follows. A PoS tagger is utilized to extract the PoS tags of every word in the texts. By converting every word into its corresponding PoS tags, we extract the PoS tag sequences from the given text in the data set. Besides PoS tags, we also used stop words to help extract information. Stop words are words that do not add contextual information to the sentences, nor do they indicate the context or meaning of a sentence. Either because they are of little significance in expressing the conception (like prepositions, conjunctions, etc.), or they are words that frequently appear in the specific speech samples that they do not contribute to the classification of machine learning models. Following the fact mentioned above, for each speech sample, 4 instances that are composed of different information (*i.e.* original texts, PoS tag sequences, and stop words list), are created. These instances are generated as described below (Fig. 3).

- **An instance with only context words**: All of the words in this case are context words. As previously noted, this is a typical method for deleting "noisy" text parts and enhances classification in a variety of natural language processing applications, including sentiment analysis [4]. Previous studies, such as [39], have used this method in the field of dementia and CI detection. The speech samples processed in this manner are used to create a data set we refer to as $\mathscr{C}$.
- **An instance without context words**: The context words in the speech samples are replaced with their PoS tags, yet we keep the stop words as they are. Although it is counterintuitive, we process the data in this way because it allows us to notice when a phrase or paragraph does not follow the natural flow of language and reveal common language patterns regardless of the context. Despite its lack of value in tasks such as sentiment analysis or hate speech detection, we believe that this information is highly useful when dealing with the issue of dementia diagnosis. The speech samples processed in this manner are used to create a data set we refer to as $\mathscr{P}$.
- **An instance where the words in the original text sample is coupled with their PoS tags**: In this case, we extracted the PoS tags of the original speech and create sequences of both. In training the network, both sequences (the text's word sequence and the PoS tag sequence) are fed into our neural networks. It is worth mentioning that the PoS tag information gets included in the word representation in some way or another, whether using a pre-trained embedding matrix or when training one from scratch. However, because of the large size of the embedding matrix and the short length of the corpus, this information may be lost in both cases. The speech

samples processed in this manner are used to create a data set we refer to as $\mathscr{O}$.

- **An instance with only PoS tags**: The instance is made up entirely of PoS tags with no contextual information in the speech. PoS tags are related to information about syntax or phrase patterns in human language. In the image description task, where the object and context are extremely well defined, dementia may be detected by assessing simply the patterns of PoS tags. The speech samples processed in this manner are used to create a data set we refer to it as $\mathscr{T}$.

C. n-GRAM LANGUAGE MODEL AND PoS TAG BASED DEMENTIA DETECTION

In their previous works [32], [36] on the two perplexities methods, they used the original transcripts with full vocabulary in their experimental setup. As previously stated, sequences made from PoS tags combined with stop words can represent the grammar and sentence patterns information while erasing all the contextual information that could indicate the actual concept in the language.

In the example shown in Fig. 3, the sentence "He is eating the cookie all by himself!" is converted into the sequence "He is VBG the NN all by PRP!" The sequence kept almost all the grammar information from the original sentence, even though the specific contextual information is hidden by the PoS tags. This sequence could help us evaluate how grammar and sentence patterns information is contributing to the two perplexities methods to detect dementia.

Under this idea, we implemented the two perplexities methods following the work of Wanekrl, *et al.* [32]. However, we do not use the transcripts with full vocabulary to evaluate the performance. Instead, the data set $\mathscr{P}$ is used in the implementation. Data set $\mathscr{P}$ in which the sequence is only composed of PoS tags and stop words. We used the tri-gram language model, which would count the frequency distribution of 3-word sequences, in the implementation to perform the two perplexities methods. The tri-gram language model is a pure statistical language model. Thus, it does not have any capability to assign a value for the sequence that does not appear in the training data set. Yet, as the Pitt Corpus is a very limited data set containing only around 500 pieces of samples with less than 100 words in each, the occurrence of unseen sequences in the test data is likely to happen. Hence, a smoothing method that avoids this issue is necessary to ensure the system works well on the test data. Additive smoothing simply assigns a fixed value to the sequences that do not appear in the training data [34]. In this experiment, we use Laplace smoothing which is a kind of additive smoothing. In implementing the two perplexities methods, we evaluate the classification accuracy by using a Leave-One-Subject-Out (LOSO) scheme. Each time, we hold all the samples of one subject out as test data and use the rest as training data. This is because, in the Pitt Corpus data set, many subjects visit the doctor more than one time, and they contribute more than one speech sample to the data set. While describing the cookie theft figure, speech samples of the same subject may use similar sentence patterns or vocabulary to describe the figure. Thus, to prevent the model from learning the identity information of the subject, we hold all samples of one subject out each time. Following the two perplexities methods of Wankerl, *et al.* [32], we train two language models on dementia data and non-dementia data respectively and use the two language models to calculate the perplexities of the test samples. By checking the perplexity difference between the two language models, we decide if the test data belongs to which category by setting a threshold. The flowchart of this method is shown in Fig. 4.

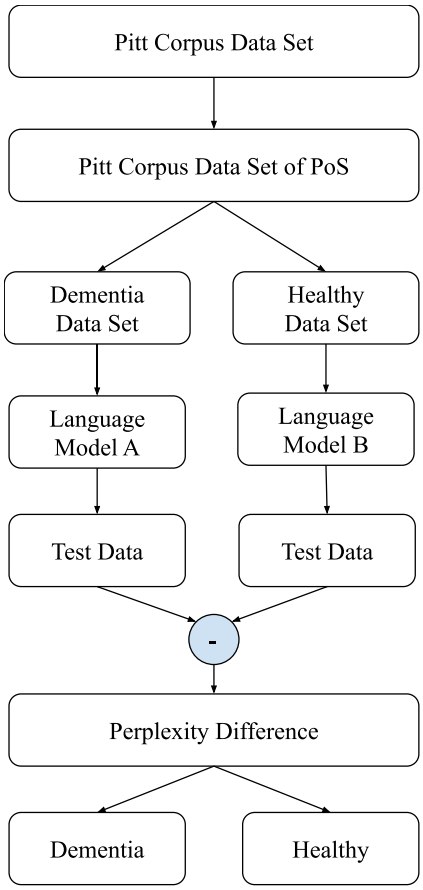

FIGURE 4. A flow chart of using two perplexities methods to detect dementia.

D. NEURAL NETWORK-BASED DEMENTIA DETECTION

We employ two neural networks for classification as previously described. The first neural network is based on the Averaged Stochastic Gradient Descent Weight-Dropped LSTM (AWD-LSTM) as proposed in [39]. The second neural network is trained from scratch using a standard attention network architecture. We will show the details of these networks in the following part.

1) PRE-TRAINED AWD-LSTM NETWORK

Original Language Model. We make use of Howard and Ruder's pre-trained language model, which is explained

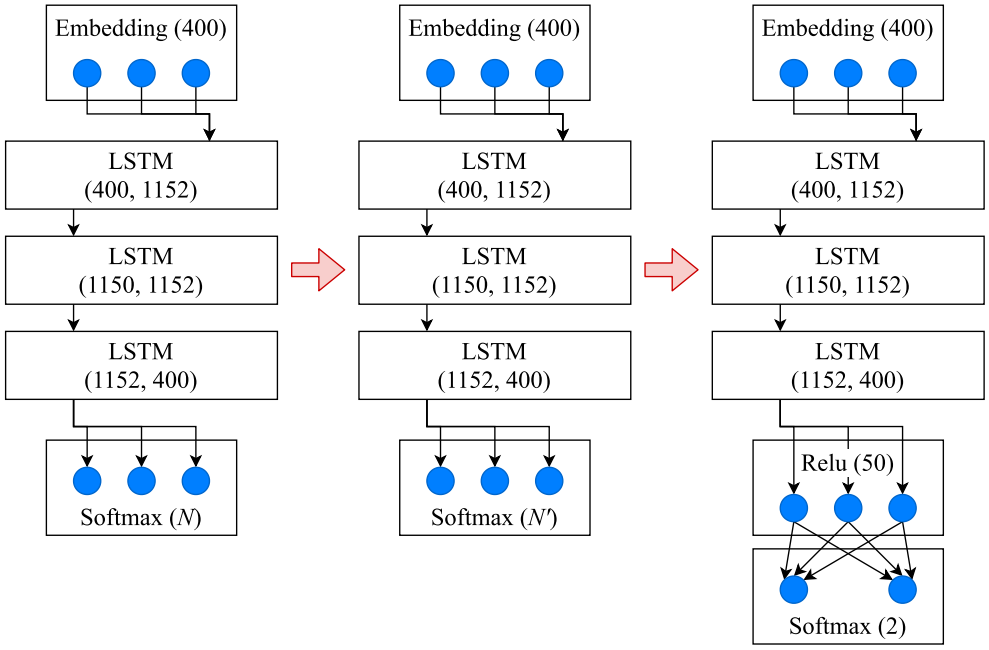

FIGURE 5. The structure of the models.

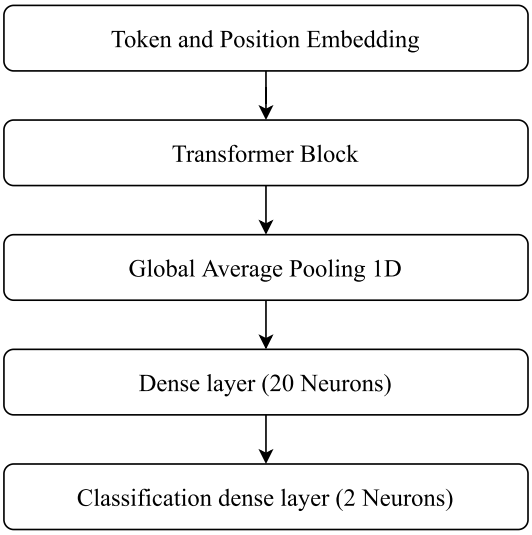

FIGURE 6. The transformer neural network architecture.

thoroughly in [39]. This language model was trained using WikiText-103 [40], a data set consisting of 28 595 pre-processed Wikipedia articles with about 103 million different words. Howard and Ruder proposed AWD-LSTM in their work. AWD-LSTM is a standard LSTM network with variably adjustable dropout hyperparameters at its core. The AWD construction is shown in the leftmost half of Fig. 5. As shown in the figure, after the embedding layer, there are three stacked LSTM layers followed by the prediction layer with a Softmax activation. The network's overall number of parameters is manageable with an embedding size of 400 and 1152 activations per LSTM layer.

Target Task Language Model Fine-Tuning. We fine-tune the pre-trained language model using the dementia data set in hand. We use all of the data in the data set. In this stage, the labels are not utilized at all because our aim at this point is for the language model to learn to understand specific linguistic characteristics. Linguistic characteristics here refer to how words are related to one another and the hidden meanings of slangs, etc. The size of the embedding matrix model is $N' \times 400$, where N' represents the number of different words in the data set, which is also the size of the first layer of the network. To fine-tune the model, the Universal Language Model Fine-Tuning (ULMFiT) technique proposed in [39] is employed, which involves progressively unfreezing and adjusting learning rates. The softmax layer is the first to be unfrozen, enabling its parameters to be fine-tuned using the learning rate of 0.1 for the first 1 epoch. The remaining layers are then unfrozen and adjusted with a learning rate of 0.001 for 5 epochs, after which we continue training by lowering the learning rate. The dropout rate for all the layers is set to 0.3, while the Adam optimizer's parameters β_1 and β_2 are set to 0.90 and 0.99, respectively.

Target Task Classifier. The last step in the language model adjustment process is the classification. We used two linear blocks, one with Rectified Linear Unit (ReLU) activation and the other one with softmax activation to substitute the last softmax layer in the original network. In other words, in this model, they are inserted after the three LSTM layers. This is because the model is no longer employed as a language model to predict the next word, but as a classification model to predict the text's class. The model is fine-tuned using slow unfreezing, discriminative learning rates, and slanted triangle learning rates. Unlike the previous phase, we utilize

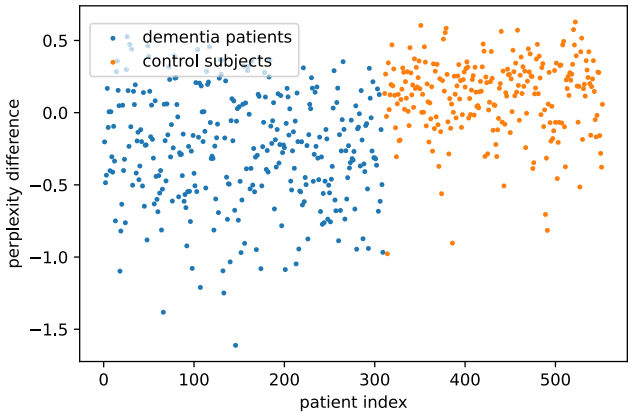

FIGURE 7. The perplexities difference between language models using only Part-of-Speech tags.

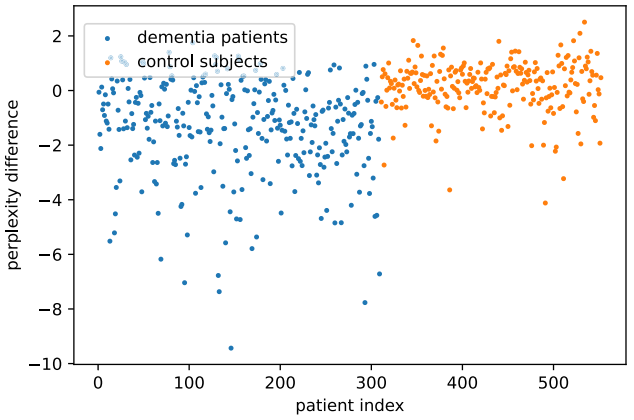

FIGURE 8. The perplexities difference between language models using Part-of-Speech tags and stop words.

the training set to fine-tune the model (while leaving the test examples to classify out) and execute the classification on the test samples left out. The Softmax and ReLU layers are unfrozen first, then fine-tuned for 1 iteration with a learning rate of 0.01. After that, we unfreeze the third LSTM besides the Softmax and ReLU layer and fine-tune the whole system to its optimal values. For both the dense layers with Softmax and ReLU activations, the learning rates are set at 0.05. The learning rate of the LSTM layer is set by following the work [39], which states that if the final layer's learning rate is η^l, prior layers should have a learning rate of $\eta^l - 1 = \eta^l/2.6$. Similarly, we unfreeze the second LSTM layer next following the same rule but with a smaller learning rate. Finally, the whole network is unfrozen and trained using the previously mentioned progressively decreasing learning rate described above.

2) ATTENTION-BASED NETWORK

We implemented a Transformer-based network architecture in the second model for classification. By replacing LSTM layers with what is known as an attention mechanism [41], the transformer network solved many issues of LSTM training difficulties. The network is made up of an embedding layer that encodes the token, its position, and segmentation information, and an attention-based transformer block, followed by 2 dense layers where the second dense layer is used for classification. The architecture of the transformer-based network is given in Fig. 6.

The main differences between the proposed methods and the previous works are summarized in the Table 1.

III. EVALUATION OF PERFORMANCE

A. PoS SEQUENCES BASED CLASSIFICATION

We first implemented the two perplexities methods using two tri-gram language models on the data set instance $\mathscr{T}$, where the sequences were made up entirely of pure PoS, tags to perform the two perplexities techniques. The perplexity difference results are presented in Fig. 7. The Y-axis of each point in the figure shows the subject's perplexity difference between the two language models. The X-axis is the patient index manually assigned to each subject. For the held-out test subject, each test subject contributes multiple samples because they visit the doctor more than one time during their treatment. There are two optional schemes for the machine learning model to perform classification. It can decide the category for each speech sample individually. It can also decide the category of the patient after concatenating the samples of him or her together as a big sample.

We determine if a test sample is a dementia patient or a healthy control subject by setting a threshold for the perplexity difference. We reached an accuracy of 75.3% in the task of classification for each patient. We achieved a classification accuracy of 71.5% in the task of classification for each sample using an equal error rate as the threshold.

We implemented the two perplexities methods using two tri-gram language models on the data set instance $\mathscr{P}$, where the context words are replaced by PoS tags and the stop words are kept as they are. The perplexity difference results are presented in Fig. 8. The range of the Y-axis is different from the previous one because we add stop words in the training process and the vocabulary for training is different, which makes the complexity of the models different.

We reached an accuracy of 80.8% in the task of classification for each patient, and 72.8% in the task of classification for each sample using an equal error rate as the threshold. The accuracy performance improved when the stop words are included, approaching that of [32]. The Receiver Operating Characteristic (ROC) curve of this experiment is also shown in Fig. 9. The ROC curve shows the true positive rate and false positive rate of our method under different thresholds. It shows the trade-off between sensitivity and specificity. We achieved an AUC of 0.78. To diagnose dementia, we utilized only 36 PoS tags with 127 stop words. It drastically lowers the cost of calculation and annotation while maintaining high accuracy.

In addition, we implemented experiments to see how much influence the high-frequency words have on the

TABLE 1. Comparison between our research and the referred research.

	ours	wankerl [32]	cohen [36]
pre-processing of the transcripts	pre-processing with PoS tagger and stop words	used the original transcripts	used the original transcripts
classifiers	n-gram language models; LSTM language models; transformer language models	n-gram language models	LSTM language models
information source	syntax information or semantic information or both	full transcript	full transcript
research methods	analysis with extracted individual information components	classification	analysis with artificially synthesized data

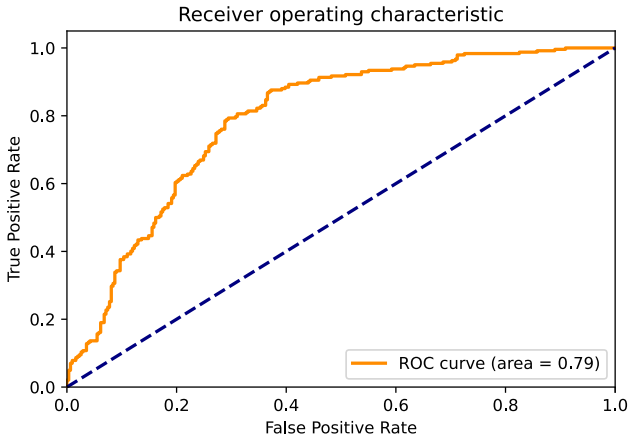

FIGURE 9. Receiver Operating Characteristic curve of classification performance for using sequences of PoS tags and stop words.

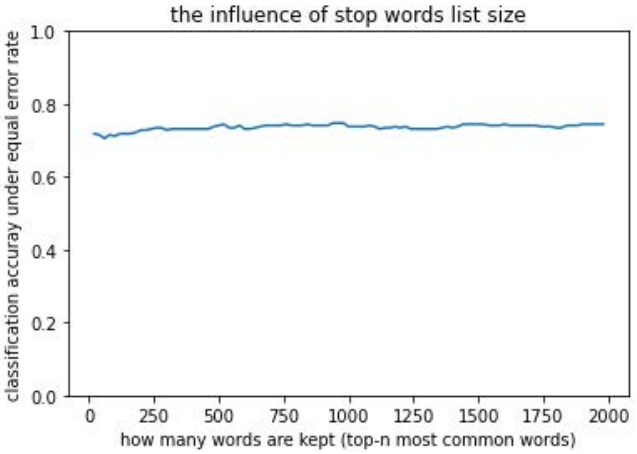

FIGURE 10. The influence of stop words list size: from 20 most used words to 2000 most used words.

two perplexities methods. In the previous experiments, we used default stop words in the Natural Language Toolkit (NLTK) [42] as the high-frequency words to keep, because they do not have any contextual information. However, despite being on top of the list of the most commonly used words in English, stop words are not necessarily the only ones on this list. In the following experiments, we created the full list of the top English words according to common corpora, where each word is ranked based on its appearance frequency. Using this list, we pick the top N words to form the actual lists of words to keep. We composed the lists that we are going to keep strictly following the order of English word frequency. In the experiments, we create 100 sub-lists following this logic. The 1st list includes the 20 most used words in English. The 2nd list includes the 40 most used words in English. The 3rd list includes the 60 most used words in English. The 100th list includes the 2000 most used words in English. For each of the lists, we re-implement the two perplexities method and see how the accuracy behaves. The results are shown in Fig. 10. When the size of kept words list is less than 260, the results demonstrate that increasing the size of the kept words list improves the classification accuracy moderately. Increasing the size of the list once it reaches beyond 260 does not improve the classification accuracy. We discovered that increasing the size of the kept words list does not always promise an increase in the performance of the n-gram language model based on the two perplexities technique. To identify dementia, we assume that the n-gram language model captures particular sentence patterns from the sequences of PoS tags. Dementia symptoms are reflected not just in vocabulary choice, but also in the syntax and sentence structures.

Furthermore, the word employed is heavily influenced by the data collection. Subjects in the Pitt Corpus data collection are requested to complete a picture description task. As a result, the number of words utilized is restricted to those contents that are visible in the image. When the subject is asked to explain what he or she is seeing, the vocabulary available to him or her has already been limited to a tiny number of terms related to the image. This implies that, regardless of how large the stop words list is, the majority of terms are not utilized in the description. Due to the small size of the Pitt Corpus data set, training an n-gram language model on PoS tags with stop words saves time and prevents over-fitting. It merely handles 36 PoS tags and a few stop words on the one hand. Converting words into their PoS tags, on the other hand, reduces the likelihood of seeing unexpected words/sequences in the test data, which helps minimize over-fitting.

B. AWD-LSTM NETWORK

We display the performance of the classification using the AWD-LSTM network in Table 3. When utilizing the original

text as is (data set $\mathscr{O}$), the classification accuracy, recall, and F1 score approach 81.54%, 83.13%, and 81.59%, respectively. When only context terms are used (data set $\mathscr{C}$), these percentages decrease to 66.54%, 66.87%, and 71.18%, respectively. This emphasizes the fact that some of the categorization information is incorporated in the full text and is lost when just context words are utilized. To demonstrate this point, we look at classification results using only stop words and replacing context words with their PoS tags. Precision, recall, and F1 score are all 76.15 %, 76.51%, and 80.38% in this example, respectively. Using the data set instance $\mathscr{P}$ shows a lower classification performance. Despite not utilizing any of the context words to train the network, the findings are rather good, leading us to assume that vocabulary richness and variety are not the only factors that may be utilized to detect dementia. The sentence's grammatical structure is a solid information source for categorization.

C. ATTENTION-BASED NETWORK

The performance of the classification using the Attention-based network on some of the instances discussed above is shown in Table 4. A similar phenomenon can be observed as can be seen in the experiment when we are using the AWD-LSTM networks. The classification accuracy, recall, and F1 score are 78.46%, 80.72%, and 82.72%, respectively, while utilizing the original texts (data set instance $\mathscr{O}$). When the stop words are removed and just context words are used (when utilizing the data set $\mathscr{C}$), these KPIs fall to 70.00%, 68.47%, and 74.51%, respectively. Finally, these KPIs reach 73.08%, 72.89%, and 77.56%, respectively, when the context words are substituted by their PoS tags (when utilizing the data set $\mathscr{P}$). This is consistent with our findings when using the AWD-LSTM network. Using the data set instance $\mathscr{O}$ gives us the highest accuracy. However, considering the situation when only context words are used (instance $\mathscr{C}$), or when only non-context words are used (instance $\mathscr{O}$), the latter gives us higher classification accuracy. This is because the sentence structure is a solid information source and it is mostly conveyed by sequences of PoS tags and stop words.

D. DISCUSSION

Syntax, phonology, morphology, semantics, and pragmatics are the four major areas of linguistics. Syntax is concerned with how words are put together to form constituents (words and sentences), and then how those constituents are placed in a certain sequence to communicate meaning. In other words, it focuses on the form of sentences and what constitutes a valid sentence. The study of how linguistic utterances and their meanings are connected, as well as how context impacts meaning, is known as semantics and pragmatics. While semantics and pragmatics have traditionally been employed extensively in NLP tasks involving text categorization, the syntactic elements that may be extracted from the transcribed texts of dementia patients may be more useful in the case of dementia diagnosis.

TABLE 2. Classification Performance of the *n*-gram Language models on the 3 Sets $\mathscr{P}$ and $\mathscr{T}$.

Set	Accuracy	Precision	Recall	F1 score
Data set $\mathscr{T}$	65.15%	63.27%	90.29%	74.40%
Data set $\mathscr{P}$	72.78%	77.32%	72.82%	75.00%

TABLE 3. Classification Performance of the AWD-LSTM Model on the 3 Sets $\mathscr{C}$, $\mathscr{P}$ and $\mathscr{O}$.

Set	Accuracy	Precision	Recall	F1 score
Data set $\mathscr{C}$	66.54%	77.62%	66.87%	71.18%
Data set $\mathscr{P}$	76.15%	84.67%	76.51%	80.38%
Data set $\mathscr{O}$	81.54%	87.34%	83.13%	85.19%

TABLE 4. Classification Performance of the Attention-Based Model on the 3 Sets $\mathscr{C}$, $\mathscr{P}$ and $\mathscr{O}$.

Set	Accuracy	Precision	Recall	F1 score
Data set $\mathscr{C}$	70.00%	81.43%	68.67%	74.51%
Data set $\mathscr{P}$	73.08%	82.88%	72.89%	77.56%
Data set $\mathscr{O}$	78.46%	84.81%	80.72%	82.72%

TABLE 5. Comparison between our proposed methods and the previous methods.

	Experiment Settings	Accuracy
Wankerl [32]	LOSO and full transcripts	77.1%
Fristch [35]	LOSO and full transcripts	85.6%
Cohen [36]	LOOCV and full transcripts	87.2%
Ours	LOSO with syntax component	72.8%
Ours	10-fold CV with syntax component	76.2%
Ours	10-fold CV with full transcripts	81.5%

In Table 2, we show that the *n*-gram language model-based two perplexities methods can detect dementia with an accuracy of 72.78% without any context words. Fig. 10 shows that keeping more high-frequency words could improve the performance of the system. Yet, after the number of kept words reaches beyond 260, keeping more words does not necessarily improve the performance. We believe that this is because, among the most used English words, words without context information contribute to the proper grammar structure account for the main part. By keeping these words in the data, more grammar information could be conveyed by the sequence of PoS tags and kept words. After the number of kept words goes beyond 260, the later included words are mostly conveying specific context information but they do not add much information in terms of grammar.

We showed in Tables 3 and 4 that maintaining only the part of texts that convey their context and subject (keeping only context words) results in a significant decline in classification performance using both classifiers. Keeping just the syntactic component of the information, on the other hand, results in a higher performance, while it is still inferior to utilizing the complete text. This indicates that a person's ability to construct grammatically accurate phrases may be used to detect dementia patients. This is consistent with our findings in the first experiment utilizing *n*-gram language models, in which we utilized perplexity applied to the syntactic section of the text to diagnose dementia using multiple language models.

Using the complete text without discarding any information, on the other hand, yields the greatest accuracy, indicating that both the syntactic and semantic components of the texts are required for better categorization.

IV. CONCLUSION

In this paper, we utilized a data set of transcribed texts obtained from dementia patients and control people to conduct dementia detection via fine-tuning applied to a common language model. Unlike previous studies, in which stop words are removed from the text, we investigated the possibility of using the stop words themselves, since they provide non-contextual information that might aid in the detection of dementia. For this, we created three neural network models: one that solely processes context words, one that stops words with patterns of PoS tag sequences, and one that combines the two. We also implemented the two perplexities methods based on *n*-gram language models with different information sources. We demonstrate that both grammar and vocabulary contribute equally to categorization via experiments: the first model achieves an accuracy of 70.00%, the second model achieves an accuracy of 76.15%, and the third model achieves an accuracy of 81.54% under 10-fold cross-validation. The *n*-gram based two perplexities methods achieve an accuracy of 72.78% under LOSO cross-validation. Our results indicate that the information encoded in the text structure and the grammatical structure of the sentences have a larger role in categorization than the context itself.

This research provided an analysis of the contribution of different language components. However, it is analyzing the topic using simple tools like PoS tags. In the future study, the sophisticated parser could be used to separate the sentences into better and finer features. Deep learning usually does not work in small data sets as smartly as in large ones. However, by analyzing and exploring the different feature representation, we could represent some hidden features by encoding them in an explicit way, which helps improve the performance of the deep learning methods. From another point of view, data augmentation has been widely used in many classification tasks. However, in dementia detection, data augmentation has not been fully explored because it is difficult to generate high-quality and meaningful text or speech due to a lack of data. However, breaking apart these sentences into lower-level representation could help to solve the problem. Compared with generating the text data, it is easier and more feasible to generate meaningful lower-level sequence data like PoS tag sequences.

REFERENCES

[1] J. A. Bourgeois, J. Seaman, and M. E. Servis, "Delirium, dementia, and amnestic and other cognitive disorders," in *The American Psychiatric Publishing Textbook of Psychiatry*. Washington, DC, USA: American Psychiatric, 2008, p. 303.

[2] S. Saxena, M. Funk, and D. Chisholm, "World health assembly adopts comprehensive mental health action plan 2013–2020," *Lancet*, vol. 381, no. 9882, pp. 1970–1971, 2013.

[3] *First WHO Ministerial Conference on Global Action Against Dementia: Meeting Report, WHO Headquarters, Geneva, Switzerland, 16–17 March 2015*, World Health Org., Geneva, Switzerland, 2015.

[4] F. Rudzicz, R. Wang, M. Begum, and A. Mihailidis, "Speech recognition in Alzheimer's disease with personal assistive robots," in *Proc. 5th Workshop Speech Lang. Process. Assistive Technol.*, 2014, pp. 20–28.

[5] M. Conway and D. O'Connor, "Social media, big data, and mental health: Current advances and ethical implications," *Current Opinion Psychol.*, vol. 9, pp. 77–82, Jun. 2016.

[6] S. Kato, H. Endo, A. Homma, T. Sakuma, and K. Watanabe, "Early detection of cognitive impairment in the elderly based on Bayesian mining using speech prosody and cerebral blood flow activation," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 5813–5816.

[7] K. Ritchie, I. Carrière, L. Su, J. T. O'Brien, S. Lovestone, K. Wells, and C. W. Ritchie, "The midlife cognitive profiles of adults at high risk of late-onset Alzheimer's disease: The PREVENT study," *Alzheimer's Dementia*, vol. 13, no. 10, pp. 1089–1097, Oct. 2017.

[8] K. Wada, T. Shibata, T. Musha, and S. Kimura, "Robot therapy for elders affected by dementia," *IEEE Eng. Med. Biol. Mag.*, vol. 27, no. 4, pp. 53–60, Jul. 2008.

[9] K. Fraser, J. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *J. Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2016.

[10] R. S. Bucks, S. Singh, J. M. Cuerden, and G. K. Wilcock, "Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance," *Aphasiology*, vol. 14, no. 1, pp. 71–91, Jan. 2000.

[11] C. Thomas, V. Keselj, N. Cercone, K. Rockwood, and E. Asp, "Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech," in *Proc. IEEE Int. Conf. Mechatronic Automat.*, vol. 3, 2005, pp. 1569–1574.

[12] D. S. Geldmacher and P. J. Whitehouse, "Evaluation of dementia," *New England J. Med.*, vol. 335, no. 5, pp. 330–336, 1996.

[13] H. C. Hendrie, "Epidemiology of dementia and Alzheimer's disease," *Amer. J. Geriatric Psychiatry*, vol. 6, no. 2, pp. S3–S18, 1998.

[14] J. T. O'Brien, "Role of imaging techniques in the diagnosis of dementia," *Brit. J. Radiol.*, vol. 80, no. 2, pp. S71–S77, Dec. 2007.

[15] K. Engedal, A. Brækhus, O. A. Andreassen, and P. H. Nakstad, "Diagnosis of dementia—Automatic quantification of brain structures," *Tidsskrift Den Norske Legeforening*, vol. 132, pp. 1747–1751, Aug. 2012.

[16] K. Lopez-de-Ipiña, J. B. Alonso, J. Solé-Casals, N. Barroso, P. Henriquez, M. Faundez-Zanuy, C. M. Travieso, M. Ecay-Torres, P. Martinez-Lage, and H. Eguiraun, "On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature," *Cognit. Comput.*, vol. 7, no. 1, pp. 44–55, 2015.

[17] K. K. Zakzanis, S. J. Graham, and Z. Campbell, "A meta-analysis of structural and functional brain imaging in dementia of the Alzheimer's type: A neuroimaging profile," *Neuropsychol. Rev.*, vol. 13, no. 1, pp. 1–18, 2003.

[18] G. Bonifacio and G. Zamboni, "Brain imaging in dementia," *Postgraduate Med. J.*, vol. 92, no. 1088, pp. 333–340, 2016.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[20] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.

[22] E. L. Campbell, L. Docío-Fernández, J. J. Raboso, and C. García-Mateo, "Alzheimer's dementia detection from audio and text modalities," 2020, *arXiv:2008.04617*.

[23] S. Luz, S. de la Fuente, and P. Albert, "A method for analysis of patient speech in dialogue for dementia detection," 2018, *arXiv:1811.09919*.

[24] B. Mirheidari, D. Blackburn, T. Walker, M. Reuber, and H. Christensen, "Dementia detection using automatic analysis of conversations," *Comput. Speech Lang.*, vol. 53, pp. 65–79, Jul. 2019.

[25] A. Garcia-Silva, C. Berrio, and J. M. Gomez-Perez, "Understanding transformers for bot detection in Twitter," 2021, *arXiv:2104.06182*.

[26] J. Li, J. Yu, Z. Ye, S. Wong, M. Mak, B. Mak, X. Liu, and H. Meng, "A comparative study of acoustic and linguistic features classification for Alzheimer's disease detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6423–6427.

[27] B. Roark, M. Mitchell, J.-P. Hosom, K. Hollingshead, and J. Kaye, "Spoken language derived measures for detecting mild cognitive impairment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2081–2090, Sep. 2011.

[28] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," in *Proc. Workshop Comput. Linguistics Clin. Psychol. From Linguistic Signal Clin. Reality*, 2014, pp. 27–37.
[29] C. Pope and B. H. Davis, "Finding a balance: The carolinas conversation collection," *Corpus Linguistics Linguistic Theory*, vol. 7, no. 1, pp. 143–161, Jan. 2011.
[30] L. Toth, I. Hoffmann, G. Gosztolya, V. Vincze, G. Szatlóczki, Z. Bánréti, M. Pákáski, and J. Kálmán, "A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech," *Current Alzheimer Res.*, vol. 15, no. 2, pp. 130–138, 2018.
[31] A. König, A. Satt, A. Sorin, R. Hoory, O. Toledo-Ronen, A. Derreumaux, V. Manera, F. Verhey, P. Aalten, P. H. Robert, and R. David, "Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease," *Alzheimer's Dementia, Diagnosis, Assessment Disease Monitor.*, vol. 1, no. 1, pp. 112–124, Mar. 2015.
[32] S. Wankerl, E. Nöth, and S. Evert, "An n-gram based approach to the automatic diagnosis of Alzheimer's disease from spoken language," in *Proc. Interspeech*, Aug. 2017, pp. 3162–3166.
[33] D. Jurasfky and J. H. Martin, *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New York, NY, USA: Pearson, 2000.
[34] T. Dunning, "Statistical identification of language," Comput. Res. Lab., New Mexico State Univ., Las Cruces, NM, USA, 1994.
[35] J. Fritsch, S. Wankerl, and E. Noth, "Automatic diagnosis of Alzheimer's disease using neural network language models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 5841–5845.
[36] T. Cohen and S. Pakhomov, "A tale of two perplexities: Sensitivity of neural language models to lexical retrieval deficits in dementia of the Alzheimer's type," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1946–1957.
[37] H. Bird, M. A. Lambon Ralph, K. Patterson, and J. R. Hodges, "The rise and fall of frequency and imageability: Noun and verb production in semantic dementia," *Brain Lang.*, vol. 73, no. 1, pp. 17–49, Jun. 2000.
[38] B. MacWhinney. (1999). *Talkbank*. [Online]. Available: http://talkbank.org
[39] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018, *arXiv:1801.06146*.
[40] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," 2017, *arXiv:1708.02182*.
[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[42] E. Loper and S. Bird, "NLTK: The natural language toolkit," 2002, *arXiv: Cs/0205028*.

CHUHENG ZHENG (Member, IEEE) received the B.E. degree from the School of Information Science and Engineering, Harbin Institute of Technology, in 2020. He is currently pursuing the master's degree with the Graduate School of Science and Technology, Keio University. His research interest includes early dementia detection based on deep learning. He is a member of IEICE.

MONDHER BOUAZIZI (Member, IEEE) received the Bachelor of Engineering (Diploma) degree in communications from SUPCOM, Carthage University, Tunisia, in 2010, and the M.E. and Ph.D. degrees from Keio University, in 2017 and 2019, respectively. He worked as a Telecommunication Engineer (access network quality and optimization) for three years with Ooredoo Tunisia (Ex. Tunisiana). He currently works as a Specially Appointed Assistant Professor with the Ohtsuki Laboratory, Faculty of Science and Technology. His research interests include machine learning, deep learning, data mining, sensors, and signal processing. He is a member of ACM and IEICE.

TOMOAKI OHTSUKI (Senior Member, IEEE) received the B.E., M.E., and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1990, 1992, and 1994, respectively.

From 1994 to 1995, he was a Postdoctoral Fellow and a Visiting Researcher in electrical engineering at Keio University. From 1993 to 1995, he was a Special Researcher of Fellowships of the Japan Society for the Promotion of Science for Japanese Junior Scientists. From 1995 to 2005, he was with the Science University of Tokyo. In 2005, he joined Keio University, where he is currently a Professor. From 1998 to 1999, he was with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley. He has published more than 215 journal articles and 415 international conference papers. He is engaged in research on wireless communications, optical communications, signal processing, and information theory. He was a recipient of the 1997 Inoue Research Award for Young Scientist, the 1997 Hiroshi Ando Memorial Young Engineering Award, the Ericsson Young Scientist Award, in 2000, the 2002 Funai Information and Science Award for Young Scientist, the IEEE First Asia-Pacific Young Researcher Award, in 2001, the Fifth International Communication Foundation (ICF) Research Award, the 2011 IEEE SPCE Outstanding Service Award, the 27th TELECOM System Technology Award, ETRI Journal's 2012 Best Reviewer Award, and the Ninth International Conference on Communications and Networking, China, in 2014 (CHINACOM'14) Best Paper Award. He served as the Chair for IEEE Communications Society, Signal Processing for Communications and Electronics Technical Committee. He served as a Technical Editor for the *IEEE Wireless Communications Magazine* and an Editor for *Physical Communications* (Elsevier). He is currently serving as an Area Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY and an Editor for the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS. He has served as the General-Co Chair, the Symposium Co-Chair, and the TPC Co-Chair for many conferences, including IEEE GLOBECOM 2008, SPC, IEEE ICC 2011, CTS, IEEE GLOBECOM 2012, SPC, IEEE ICC 2020, SPC, IEEE APWCS, IEEE SPAWC, and IEEE VTC. He gave tutorials and keynote speeches at many international conferences, including IEEE VTC, IEEE PIMRC, and IEEE WCNC. He was the Vice President and the President of the Communications Society of the IEICE. He is a Distinguished Lecturer of the IEEE, a fellow of the IEICE, and a member of the Engineering Academy of Japan.