Could Discourse Analysis Reveal Pre-frailty in Older Adults?

Manuel Abbas
Univ Rennes, Inserm
LTSI - UMR 1099
F-35000 Rennes, France
manuel.abbas@univ-rennes.fr

Joaquim Prud'Homm
Univ Rennes, Inserm
LTSI - UMR 1099
F-35000 Rennes, France
joaquim.prud-homm@inserm.fr

Fabien Lemoine
Univ Rennes, Inserm
LTSI - UMR 1099
F-35000 Rennes, France
fabien.lemoine@univ-rennes.fr

Guy Carrault
Univ Rennes, Inserm
LTSI - UMR 1099
F-35000 Rennes, France
guy.carrault@univ-rennes.fr

Dominique Somme
Univ Rennes, CHU Rennes, CNRS
ARENES - UMR 6051, RSMS - U 1309
F-35000 Rennes, France
dominique.somme@chu-rennes.fr

Régine Le Bouquin Jeannès
Univ Rennes, Inserm
LTSI - UMR 1099
F-35000 Rennes, France
regine.le-bouquin-jeannes@univ-rennes.fr

***Abstract*—The promotion of healthy aging is one of the main challenges of our society in the face of the aging population phenomenon. Early detection of frailty, which is a geriatric syndrome that alters the lifestyle of the elderly and threatens their desire to live independently, is a challenge, and to date there is no consensual method for this detection. In this paper, we analyze the transcribed discourse of older adults, in response to a series of open-ended questions. To this end, the answers of 34 subjects aged over 80 years were transcribed into texts. The objective of this study is to evaluate the ability of discourse analysis to separate robust from pre-frail individuals. Explicitly, we ran through two text classification approaches: (i) the application of a linear model on the TF-IDF technique and (ii) the application of a recurrent neural network after encoding the words with the Word2Vec algorithm using pre-learned information. Our models identify potentially informative words in the corpus of older individuals recruited in this study. This study suggests that text analysis provides promising diagnostic utility in the analysis of frailty, and opens the door for further research.**

***Index Terms*—frailty, older adults, transcribed discourse, text analysis, classification**

I. Introduction

As citizens live longer and healthier lives, new solutions should emerge to promote healthy aging [1]. Frailty is a geriatric syndrome defined as "a state of increased vulnerability to poor resolution of homeostasis after stress" [1]. Being frail increases the risk of falls, disability, hospitalization, nursing home admission, and death at 2 years compared with robust individuals [1], [2]. Frailty is profoundly related to physical, psychological and cognitive impairment [2], [3]. In some ways, it also describes "a person's overall resilience and how it relates to his or her chances of recovering quickly from a health problem" [4]. Therefore, early detection of frailty, coupled with person-centered support, could help older adults stay healthy and live independently as long as possible. There are more than 50 frailty assessment tools [5], with frailty phenotype being the best known [6]. It characterizes frailty following five criteria, namely (i) unintentional weight loss, (ii) self-reported exhaustion, (iii) weakness , (iv) slow walking speed and (v) low physical activity. If no criteria are met, the subject is said to be robust, and if 1 to 2 criteria are met, he/she is qualified as pre-frail. Older adults with more than 2 criteria are frail. Although numerous conceptual models exist for frailty screening, the majority focuses on the physical function and activity metrics (such as gait speed), and only 27 tools explore psychological or social aspects related to frailty [5]. Moreover, spontaneous discourse is not exploited in those tools. However, Garcia *et al.* have shown that linguistic parameters might be good indicators to cognitive impairment and frailty [7]. Expressly, older adults were asked to provide their perspective on the concept of healthy aging, and their verbal responses underwent content analysis. The number of ideas expressed in their answers seemed to be associated to a more positive prognosis in mental and physical health. Thus, we hypothesize that discourse analysis could be a novel and relevant solution for the early detection of frailty, and constitute another dimension than the physical activity.

Natural language processing (NLP) is the use of computer techniques to automatically analyze and represent human language [8]. Research on NLP has accelerated text analysis and allowed machines to carry out a wide variety of natural language related tasks. NLP is currently a trending research topic [9], [10] and has provided state-of-the-art results in several fields like sentiment analysis, language modeling, translation tasks, and clinical research to name a few [11]–[13]. Nevertheless, to our knowledge, text analysis in the context of frailty has been infrequently addressed in the literature. It is worth mentioning that researchers from the University of Patras have investigated the written texts of older adults to predict their frailty level [14]. They started by an automatic

This work was supported by the French National Research Agency (ANR) under Grant ANR-17-CE19-0024-01, by Fondation de l'Avenir in the context of EPOPEE Project under grant AP-RM-21-006, and by the health technologies hospital-university federation (FHU TECH-SAN).

extraction of 160 features, such as the existence of keywords, sentiment value (sad to happy) and other variables carrying statistical properties of the older person's text. The most discriminant subset of features was chosen using Pearson's correlation coefficient. Finally, a majority voting technique was employed, combining the best performing models in an ensemble classifier. They achieved an accuracy of 64% by classifying older individuals into three groups, namely robust, pre-frail, and frail. Although this study is interesting, it shows some limitations. Firstly, an open discussion instead of simple/closed-ended questions with the subject is required to extract as many elements as possible. Secondly, the most relevant spoken words for frailty classification are not sufficiently discussed. In addition, the most sensitive topics provoking unsolicited responses and leading to early detection of frailty are still unknown.

Consequently, the aim of the present study is to examine the capacity of older people's discourse to reveal pre-frailty or the start of the physical impairment process. We collected answers from community-dwelling older adults to 8 open-ended questions and transcribed them into texts, then examined two approaches. On the one hand, we assigned a weight to each word using a statistical measure called Term Frequency - Inverse Document Frequency (TF-IDF) to vectorize their speech. The resulting vectors were fed into a logistic regression to classify older adults. On the other hand, we tested another approach by applying a word embedding method called Word2Vec, which vectorizes the words instead of assigning a simple weight. This transformation was followed by the application of a deep learning classifier to identify the level of frailty.

The paper is organized as follows. Section II describes data collection and pre-processing, and introduces the approaches to investigate the relationship between the spontaneous speech and the frailty status. Section III illustrates and interprets the experimental results of this study. Section IV discusses the limitations of this study while section V concludes the paper and proposes future work.

II. Materials and Methods

A. Data Preparation

Thirty-four community-dwelling older adults were recruited, and they were offered a two-year health follow-up. It is worth mentioning that the participants had to be autonomous and not classified as "frail" according to the frailty phenotype to be included in the study. They were visited three times during this period by a specialist in psychology: the first visit taking place at the start of the study, the second visit taking place a month later, and the third one after more than 6 months. The psychologist asked them eight open questions during his visits that are mentioned in Table I. No specific answers were expected, and the subjects did not prepare their answers in advance. It was a sort of an open discussion between the older adult and the psychology specialist. For instance, the following two replies were obtained after asking the first question Q1: (i) "*I'm fine, let's say not too bad overall. I'm not complaining.*"; (ii) "*My morale is bad ... It is very bad right now*". The spontaneous speech of the older individuals was recorded using a smartphone, and a verbatim transcription was written after the visit, so that the exact spoken words and sentences were saved as texts. A geriatrician visited the subjects every three months simultaneously, to assess their frailty condition using the frailty phenotype (FrP) [6]. Two populations were identified, namely (a) robust people and (b) pre-frail people. Approval of all ethical, protocols, and use of data was granted by the Ouest VI Institutional Review Board of Morvan University Hospital of Brest, France under Approval No. 1428 (IDRCB: 2019-A02316-51, RIPH: 21.02302.000026).

TABLE I
The open questions which were asked during the interviews

ID	Question
Q1	How is your morale?
Q2	What do people around you think about your health?
Q3	What do people around you think about your morale?
Q4	What are your habits in terms of health monitoring?
Q5	What are your reasons for participating in the study?
Q6	What does aging mean to you?
Q7	What are the advantages and disadvantages of aging?
Q8	How old do you feel you are most of the time?

To create our dataset, which is going to be considered as the corpus of older individuals, the transcribed discourses were cleaned and segmented, then labeled using the score of FrP. Each sample of the corresponding dataset was labeled as 0 if the subject is robust or 1 if he/she is pre-frail. Another label was assigned to each sample, indicating the ID of the question. It is worth noting that 52.7% of data belonged to robust subjects and 47.3% to pre-frail ones.

B. Text Analysis and Classification

The goal of this paper is to investigate the relevance or the prominence of the expressed words in the older adults' answers in the context of frailty screening, in order to identify any potential correlation. Therefore, the words were encoded into numerical data to feed machine learning classifiers using two approaches. Firstly, we used a classical machine learning technique by employing a statistical measure on texts and then by applying a linear model for classification. Secondly, another word embedding method was considered to encode data using pre-learned information, before proceeding to the classification using a deep learning model, increasing the computational complexity.

1) TF-IDF: the Term Frequency–Inverse Document Frequency (TF-IDF) technique was first used to determine the mathematical significance of words [15]. It allows us to quantify the importance or relevance of string representations in a sample in the dataset. TF-IDF can be broken down into two parts, namely (i) the term frequency (TF), which is an indicator of how often a term occurs in a sample, and (ii) the inverse document frequency (IDF), which is an indicator of the relative rarity of a term in the dataset. Expressly, TF works by examining the frequency of a particular term or word relative to the sample, and IDF indicates how common (or not) a word is in the corpus. The final TF-IDF value is obtained by multiplying these two aforementioned values together. Thus, according to TF-IDF, the importance of a term is inversely related to its frequency in the samples. The higher the TF-IDF score, the more important or relevant the term is. If a term becomes less relevant, its TF-IDF score approaches 0. A value equal to zero indicates the absence of the term in the sample. Consequently, to vectorize the texts, we began by removing stopwords and commonly used words such as "with", "this", "in", "he", "she", the verbs "have" and "be" to name a few, using the Natural Language Toolkit NLTK (version 3.8.1) of Python (version 3.9.16). Lemmatisation was then applied, grouping together the inflected forms of a word to its base root mode. For instance, the answer "*My morale is bad*" becomes "*morale bad*" after this transformation. As a result, a total of 1853 words were found in the whole dataset (the older adults corpus). The TF-IDF method was then applied to assign a weight for each term in the sample. Explicitly, each sample became a vector of size 1853, where each element represented a word in the corpus. For example, the response "*morale bad*" became a vector of zeros, except its 160$^{\text{th}}$ element that is equal to 0.77 (weight of the word "bad") and its 1020$^{\text{th}}$ element that is equal to 0.62 (weight of the word "morale"). Once the samples were vectorized, a logistic regression was applied to classify the responses into two groups.

2) Word2Vec: the Word2Vec method was also used in our analysis. This technique scans an entire corpus and creates a vector for each word [16], positioning the word in a N-dimensional space, contrarily to TF-IDF which represents each word by a scalar. It accounts for semantic similarities in a language by capturing the relationships between the words. It relies on the distributional hypothesis [17]: the words which appear in the same context have similar meanings. In Word2Vec, an unsupervised learning process is performed using artificial neural networks to create a model that generates word vectors. Seeing the limited size of our dataset, we did not perform the aforementioned training process to create the word representation model, but we downloaded an existing representation from the open-source library gensim, namely "Word2Vec-google-news-300" (WGN-300). This model contains 300-dimensional vectors for 3 million words and phrases. Explicitly, we started by initializing the tokenizer (splitting raw texts into tokens) and we fitted it to the training data, so that each word was represented by an index. We then mapped the aforementioned pre-learned model to our tokenization and used WGN-300 to build weights of the embedding matrix, which took the integer-encoded vocabulary and looked up the embedding vector for each word-index. The following deep learning model (5 layers) was applied for classification:

- Embedding layer: using the pre-trained weights
- Gated Recurrent Unit (GRU) layer: memory units $\rightarrow$ 15
- Dropout layer: rate $\rightarrow$ 0.2 [helps prevent overfitting]
- Dense layer: units $\rightarrow$ 64 ; activation function $\rightarrow$ 'relu'
- Dense layer: units $\rightarrow$ 2 ; activation function $\rightarrow$ 'softmax' [output layer]

III. Results and Discussion

To achieve the objective of this study, *i.e.* investigate the ability of discourse analysis to distinguish between robust and pre-frail older adults, the binary classification was done using the dataset described in section II-A, with the pre-frail population being the positive class. Leave-Subject-Out (LSO) cross-validation was applied to evaluate the performance of both TF-IDF and Word2Vec. At each iteration i, the model was fitted to the data of 33 subjects then tested on the remaining one. This process resulted in 34 values of accuracy δ_i, each one representing the ability of the model to correctly identify the status of the i^{th} subject. The experimental results are illustrated in Table II, in terms of the minimum (Min) and maximum (Max) values of δ, the average cross-validation accuracy, as well as the sensitivity (Sen) and the specificity (Spe) using true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN):

$$\begin{cases} \text{Accuracy} = \frac{1}{34}\sum_{i=1}^{34}\delta_i \\ \text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \\ \text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \end{cases}$$

The 34 values achieved by TF-IDF are distributed between 25% and 76.47%, with an average accuracy of 53.83%. As mentioned previously, a logistic regression model was used for the text classification. To calculate the output y in order to make a prediction, this method attributes a weight ω_j for each word X_j in the sample then adds a bias β before applying the Sigmoid function:

$$y = \frac{1}{1 + exp[-(\beta + \sum_{j=1}^{1853}\omega_j X_j)]} \tag{1}$$

The advantage of this linear model relies in the results interpretation. The weights of this model can be sorted in order to extract the most important words in the process of frailty identification, since the parameters correspond indirectly to the expressed words. According to TF-IDF, the words that count most in this context (or the words with the highest coefficients in this classification process) are "*trouble*", "*easy*", "*body*", "*kind*", "*husband*", "*son*", "*alone*", "*old*", "*fall*", "*difficulty*", "*pain*", "*daughter*", "*sleep*". When it comes to Word2Vec, this method achieved better results. The values of δ_i range between

TABLE II
Classification results per subject using the leave-subject-out (LSO) cross-validation

Method	**Metrics (%)**				
	Min	**Max**	**Accuracy**†	**Sen**‡	**Spe**‡
TF-IDF	25	76.5	53.83 ($\pm$ 13.6)	43.4	61.4
Word2Vec	44.7	92.6	67.64 ($\pm$ 12.3)	54.4	76.2

† the mean of the 34 accuracy values ($\pm$ their standard deviation)
‡ global measurement

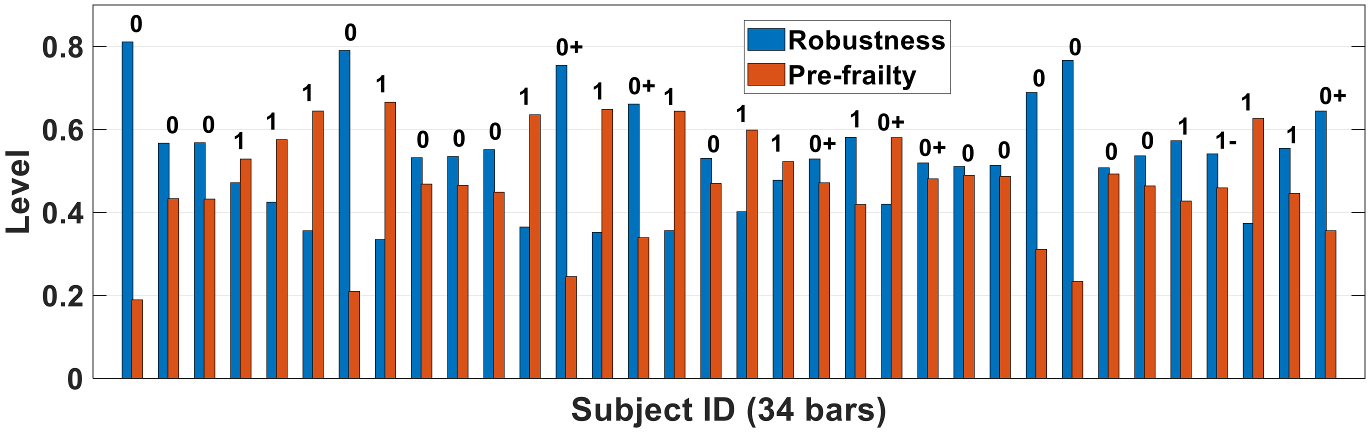

Fig. 1. The frailty cursor according to Word2Vec in the form of a probability value pair $\{z\ ,\ 1\text{-}z\}$: the blue bars represent the probability of being robust (robustness level) and the orange ones represent the probability of being pre-frail (pre-frailty level). The sum of the two bars is equal to 1. The values $\{$0, 0+, 1-, 1$\}$ represent the ground truth for each subject.

44.7% and 92.6%, with an average accuracy of 67.64%. This model seems specific (a specificity of 76.2% compared to a sensitivity of 54.4%), which means that the number of FP is somewhat low. Here, the discussed deep learning model provides its prediction in the form of a probability value pair $\{z\ ,\ 1\text{-}z\}$, z being the probability of belonging to the positive class (being pre-frail). Hence, we obtained a matrix of N rows and 2 columns for each subject, where each row represents one of his/her responses and the columns contain the values of z and $1-z$. Consequently, we calculated the median of each column for each individual, resulting in two values per subject, and we plotted the results as bars in Fig. 1 to constitute the frailty cursor. Explicitly, the blue bars represent the probability of being robust and the orange ones represent the probability of being pre-frail. This figure also shows the ground truth, with '0' being assigned if the subject is robust, '1' if he/she is pre-frail, '0+' if he/she went from robustness to pre-frailty during the study, and '1-' if he/she was pre-frail at the start and became robust at the end of the data collection. Moreover, the accuracy achieved by each question (see Table I) was calculated for each method (TF-IDF *vs* Word2Vec) and illustrated in Fig. 2. From TF-IDF point of view, the most discriminating question is that of the morale (Q1) with an accuracy of 63.05%, followed by Q5 which is related to the reason behind the participation in the study (59.22%). When it comes to Word2Vec, the accuracy of all questions exceeds 60%, with Q8 (how old they feel they are) being the top performer with a 72.58% accuracy.

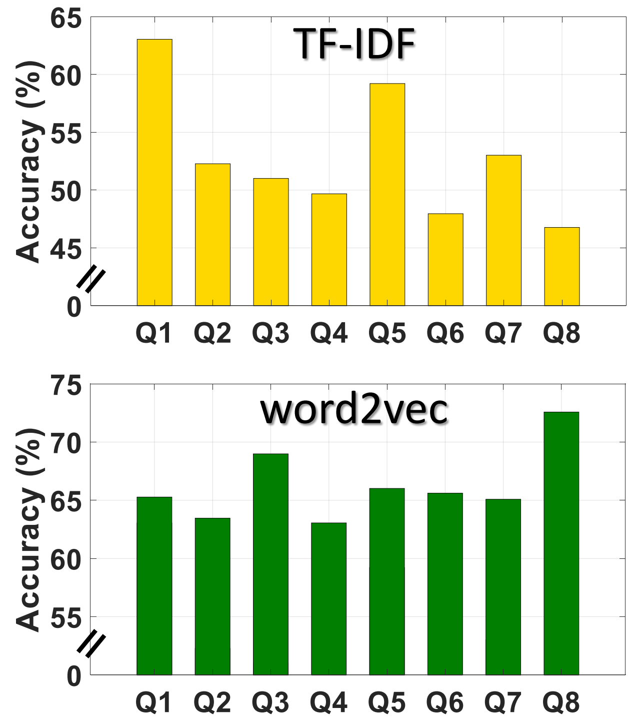

Fig. 2. The accuracy (%) achieved per question (Q1 to Q8, see Table I) following both methods, *i.e.* TF-IDF and Word2Vec.

The experimental results unveil several elements regarding the frailty topic. First, the frequency of the words expressed, which is a basic attribute of TF-IDF, provides limited information. Indeed, the accuracy for some subjects is quite low (25%), which means that this method failed to identify the status for 75% of their data. Nevertheless, TF-IDF recognized the top words, which is still unknown in this context. For instance, talking about family members ("*husband*", "*son*", "*daughter*") and physical conditions ("*body*", *trouble*", "*old*", "*fall*", "*difficulty*", "*pain*", "*sleep*") seems informative when it comes to frailty screening. Secondly, Word2Vec shows that a relationship exists between the words expressed by the older individuals and their frailty status. It was needed in our task since it captures semantic and syntactic similarity, as well as the relationship with other words. In fact, an accuracy of 67.64% based on transcribed discourse alone is quite satisfactory, given that these data are subjective and frailty detection requires much more than self-reported responses to open-ended questions. In addition, the two populations are close to each other since all participants are autonomous and live independently. Detecting pre-frailty, which is the intermediary class or the start of the frailty process, is difficult as seen in previous studies [14]. Thirdly, this study shows to what extent our models were capable to model the target variable given the answers to the eight questions (Table I). For instance, the answers to Q4, concerning habits in terms of health monitoring, does not seem informative for separating the two groups based on this corpus, unlike Q1, Q3 and Q8.

IV. Study Limitations

Although these preliminary results are quite promising and pave the way for a deeper analysis, this study presents some limitations. The TF-IDF method is an easy to compute statistical measurement. It is able to extract the most descriptive terms in a discourse, and thus it is useful as a lexical level feature. Nonetheless, it cannot capture semantics and does not focus on the context in which the words are employed. Moreover, a frequent word in a corpus is not necessarily less important than other words. The Word2Vec has a powerful architecture and can understand the meaning of words. Nevertheless, it is unable to handle out-of-vocabulary words. For instance, some words in the corpus (mainly digits and numbers which are mostly mentioned in answers to Q8) are unknown to the pre-trained model WGN-300. Moreover, handling morphologically similar words is another challenge. The homonym "bank" is a famous example. Focusing on transformer-based neural networks could be interesting and more useful in our case, BERT being an example [18].

Furthermore, the older adults were labeled using the score of Fried's frailty phenotype [6]. This phenotype mainly focuses on physical worsening, but other factors like cognitive impairment, dementia, depression, and anxiety to name a few could be important indicators to frailty. Consequently, some other scores should be evaluated to investigate the correlation between speech and these other aspects of frailty. Tests like the frailty cumulative index provides other measures that are worth evaluating [19].

V. Conclusion

This paper investigated the ability of transcribed discourse to identify pre-frailty in old-age. Two approaches were applied, namely TF-IDF and Word2Vec. This study showed that a certain correlation exists between the words expressed by the older individuals and their frailty status. More importantly, it brings a predictive perspective to the frailty screening through text analysis of older adults, and it allows the validation of the questions that must be asked to obtain a meaningful answer. This paper is a first step towards automatic detection of frailty based on spontaneous discourse. The experimental results showed that the transcribed texts separated the two populations with an accuracy of 67.64%, which is quite satisfactory for this assessment. In addition, given our dataset, the expressed words that matter most in this classification task were also revealed in this paper.

In a future work, more appropriate topic modelling algorithms like Latent Dirichlet Allocation could be performed to investigate and identify the relevant terms in this context. Moreover, other classification techniques might be examined: some relatively simple approaches such as recurrent neural networks with attention that might improve downstream classification performance, and/or more complex methods like fine-tuning BERT model or applying other transformer-based architectures. Furthermore, studying the different aspects of frailty by using the scores of different questionnaires is also going to be a future research topic.

References

[1] A. Clegg, J. Young, S. Iliffe, M.O. Rikkert, K. Rockwood, "Frailty in elderly people", The Lancet, no. 381, pp. 752-762, 2013.

[2] S. Vermeiren *et al.*, "Frailty and the Prediction of Negative Health Outcomes: A Meta-Analysis," J Am Med Dir Assoc., vol. 17, no. 12, 1163.e1-1163.e17, 2016.

[3] E.O Hoogendijk *et al.*, "Frailty: implications for clinical practice and public health", The Lancet, vol. 394, no. 10206, pp. 1365–1375, 2019.

[4] Age UK, "Understanding frailty," 2020, available: https://www.ageuk.org.uk/our-impact/policy-research/frailty-in-older-people/understanding-frailty/.

[5] J.W. Faller *et al.*, "Instruments for the detection of frailty syndrome in older adults: A systematic review," PLoS One, vol. 14, no. 4, e0216166, 2019.

[6] L. Fried *et al.*, "Frailty in older adults: Evidence for a phenotype," Journal of Gerontology: Medical Sciences, vol. 56A, no. 3, pp. 146-156, 2001.

[7] T.F. Meira Garcia *et al.*, "Number of ideas in spontaneous speech predicts cognitive impairment and frailty in community-dwelling older adults nine years later," Aging & Mental Health, vol. 26, no. 10, pp. 2022-2030, 2022.

[8] T. Young, D. Hazarika, S. Poria and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]," IEEE Computational Intelligence Magazine, vol. 13, no. 3, pp. 55-75, 2018.

[9] H. Sakai and H. Iiduka, "Riemannian Adaptive Optimization Algorithm and its Application to Natural Language Processing," IEEE Transactions on Cybernetics, vol. 52, no. 8, pp. 7328-7339, 2022.

[10] M. Omar, S. Choi, D. Nyang and D. Mohaisen, "Robust Natural Language Processing: Recent Advances, Challenges, and Future Directions," IEEE Access, vol. 10, pp. 86038-86056, 2022.

[11] R. Yang, "Machine Learning and Deep Learning for Sentiment Analysis Over Students' Reviews: An Overview Study," Preprints, 2021020108, 2021, available: https://www.preprints.org/manuscript/202102.0108/v1

[12] I. Sutskever, O. Vinyals, and Q.V. Le, "Sequence to sequence learning with neural networks," Proc. Advances Neural Information Processing Systems, pp. 3104-3112, 2014.

[13] E. H. Houssein, R. E. Mohamed and A. A. Ali, "Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review," IEEE Access, vol. 9, pp. 140628-140653, 2021.

[14] FrailSafe, "Frailty detection from text analysis," 2018, available: https://frailsafe-project.eu/news/80-text-analysis.

[15] A. Aizawa, "An information-theoretic perspective of tf–idf measures," Information Processing & Management, vol. 39, Issue 1, pp. 45–65, 2003.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean "Efficient Estimation of Word Representations in Vector Space," arXiv:1301.3781, 2013, available: https://arxiv.org/abs/1301.3781.

[17] Z.S. Harris, "Distributional Structure," WORD, vol. 10, no. 2-3, pp. 146–162, 1954.

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, 2019, available: https://arxiv.org/abs/1810.04805.

[19] K. Rockwood, A. Andrew, and A. Mitnitski, "A Comparison of Two Approaches to Measuring Frailty in Elderly People," J Gerontol A Biol Sci Med Sci, vol. 62, no. 7, pp. 738-–743, 2007.