INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Tackling the ADRESSO Challenge 2021: The MUET-RMIT System for Alzheimer's Dementia Recognition from Spontaneous Speech

Zafi Sherhan Syed[1], *Muhammad Shehram Shah Syed*[2], *Margaret Lech*[2], *Elena Pirogova*[2]
[1]Mehran University, Pakistan
[2]RMIT University, Australia
zafisherhan.shah@faculty.muet.edu.pk
muhammad.shehram.shah.syed|margaret.lech|elena.pirogova@rmit.edu.au

Abstract

This paper addresses the Interspeech Alzheimer's Dementia Recognition through Spontaneous Speech only (ADReSSo) challenge 2021. The objective of our study is to propose the approach to a three task automated screening that will aid in distinguishing between healthy individuals and subjects with dementia. The first task is to differentiate between speech recordings from individuals with dementia. The second task requires participants to estimate the Mini-Mental State Examination (MMSE) score based on an individual's speech. The third task requires participants to leverage speech recordings to identify whether individuals have suffered from cognitive decline. Here, we propose a system based on functionals of deep textual embeddings with special preprocessing steps integrating the effect of silence segments. We report that the developed system outperforms the challenge baseline for all three tasks. For Task 1, we achieve an accuracy of 84.51% compared to the baseline of 77.46%, for Task 2, we achieve a root-mean-square-error (RMSE) of 4.35 compared to the baseline of 5.28, and for Task 3, we achieve an average-f1score of 73.80% compared to the baseline of 66.67%. These results are a testament of the effectiveness of our proposed system.

Index Terms: alzheimer's dementia, computational paralinguistics, social signal processing

1. Introduction

Alzheimer's disease is a chronic neurodegenerative disorder that detrimentally impacts cognitive and physical well-being of a person. According to the World Health Organization (WHO) [1], dementia currently affects more than 50 million people worldwide, with millions of new patients being diagnosed every year.

The ever-growing use of artificial intelligence (AI) in healthcare-related applications has facilitated development of innovative and advanced medical diagnostic approaches to various types of disorders [2, 3, 4, 5]. The main advantage of such techniques is that they can be successfully employed for objective diagnosis of disorders. The limited human interference assists in reducing human errors and bias. Considerable effort has been directed towards the development of diagnostic methods which can be used to identify individuals with Alzheimer's dementia [6].

The Interspeech Alzheimer's Dementia Recognition through Spontaneous Speech only (ADReSSo) challenge 2021 [7] aims to provide a common platform to researchers to not only propose methods for automated screening of Alzheimer's dementia but also encourages researchers to compete and evaluate their work against their peers. The challenge this year may be considered as an extension of last year's ADReSS 2020 challenge [8] with an important difference. Whereas last year, the dataset contained manually transcribed and CLAN [9] annotated transcripts, this year's challenge expects participants to work with automatically generated speech transcripts.

In last year's ADReSS challenge, our developed system performed very well, achieving an accuracy of 85.42% compared to the challenge baseline of 77.00% and an RMSE score of 4.30 compared to the baseline of 5.20 for the test partition. In essence, first we observed that features derived from textual modality offer much better performance than those from the audio modality. Secondly, we had demonstrated the prowess of a simple but very effective method for representing speech transcripts of subjects as feature vectors. To that end, we had first computed deep textual embeddings (DTE) from transformer based models and applied functionals of descriptive statistics to pool their values into a feature vector.

This paper describes our proposed system [1] for tackling the ADReSSo challenge 2021. Here, we demonstrate the efficacy of a system based on functionals of deep textual embeddings with special preprocessing steps integrating the effect of silence segments. We also show the potential benefits of class-imbalance aware multi-model fusion.

2. Dataset

The ADReSSo challenge consists of two distinct datasets. The first dataset is called 'Diagnosis' and is used for Task 1 and Task 2 of the challenge. In Task 1, the objective is to differentiate between speech recordings from individuals with dementia amongst a set of recordings from healthy individuals. In Task 2, the objective is to estimate the Mini-Mental State Examination (MMSE) score based on an individual's speech. The second dataset 'Progression' is used for Task 3 of the challenge. Here, the objective is to identify, based on characteristics of their speech, whether subjects have suffered from a cognitive decline over years. For further details regarding the dataset, we refer the reader to the ADReSSo challenge baseline paper [7].

3. Methodology

A block diagram representation of our proposed system for the ADReSSo challenge is provided in Figure 1 where it can be seen that the system starts with automated speech recognition (ASR) for speech-to-text conversion. Next, we experiment with various preprocessing methods (detailed in Section 3.2). This is followed by a process of generating feature representations for transcripts using DTE as well as handcrafted features. We used handcrafted features to compare the performance of deep textual embeddings against domain-knowledge features. The

[1]Our previous work [10] provides the necessary context to our current work

http://dx.doi.org/10.21437/Interspeech.2021-1572

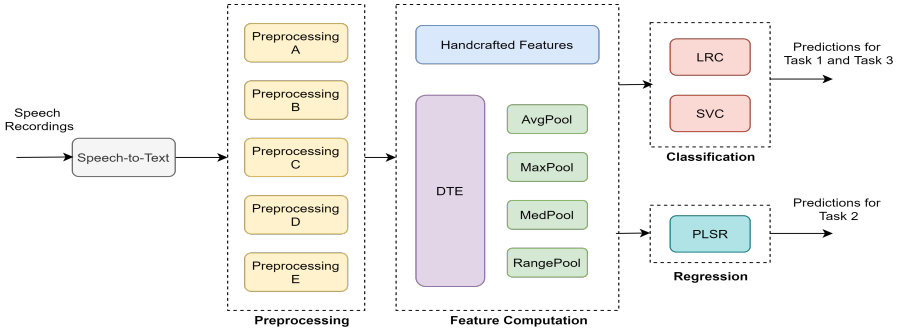

Figure 1: *Block diagram for our proposed automated screening system*

final step is to train classification and regression methods on the training partition and make predictions for the test partition.

3.1. Generating Speech Transcripts

An important aspect of the ADReSSo challenge 2021 is for participants to work without manually annotated transcripts. The dataset contains time-stamps that could be used to identify speech segments from the subject and the interlocutor. However, through preliminary experiments we discovered that these time-stamps are not always aligned with speech recordings. Although it was possible to use speaker diarization to identify segments of speech that belong only to a particular subject, we decided to use ASR to generate transcripts for the entire speech recording without diarization. We assume that (a) speech from the subject will dominate the recording and the contribution from the interlocutor will be relatively small, and (b) the speech and language from the interlocutor will also reflect the cognitive state of the subject. For example, the interlocutor will use simplified language to communicate with an individual that suffers from language impairments due to dementia.

For ASR, we experimented with wav2vec2 model from Huggingface toolkit [11], Silero [12] toolkit, and Microsoft Stream [2]. Initial results showed that the latter provided the most promising results in terms of word error rate. We plan to compare the performance of various automated ASR approaches for the task at hand in due course.

3.2. Preprocessing

We have experimented with five types of text preprocessing methods to investigate whether a particular method leads to an improvement in classification or regression tasks of the challenge.

- Preprocessing-A: Here, we resolved word contractions, removed punctuation and extra whitespaces from transcripts, and converted all text in lower-case. Thus, the entire transcript was set up as a single sentence.
- Preprocessing-B: We removed extra whitespaces and represented text in lower-case only.
- Preprocessing-C: Here, we decided to add special keywords into the speech transcript depending on the duration of a silence segment between two successive utterances. We were inspired to explore this method given the success reported by [13] using a similar technique. However, unlike Yuan et al., we test this method on pretrained embeddings only. Therefore in Preprocessing-C, we started with the setup of Preprocessing-A and if the silence duration was determined to be between 2 and 4 seconds, we added the text 'uhm'. If the silence duration was between 4 and 6 seconds, we added the word 'uhm uhm'. Finally, if the silence duration exceeded 6 seconds, we added the text 'long silence'.
- Preprocessing-D: In this method, we followed the procedure of adding special keywords as in Preprocessing-C after removing extra whitespaces and also converted text in lower-case.
- Preprocessing-E: Here, the preprocessing was performed as in Preprocessing-D except that the replacement text for silence segments was a period symbol ('. ') instead of 'uhm'. For example, if the silence duration was between 2 and 4 seconds, we added a '. ' as text. If the silence duration was between 4 and 6 seconds, we added '. . ' , and in case of the silence duration exceeding 6 seconds, we added the text 'long silence'.

3.3. Feature Computation

Once speech transcripts were generated and preprocessed, the next step was to compute textual features. As mentioned previously, we computed both, handcrafted features and deep textual embeddings. For handcrafted features, we computed a set of textual features inspired by the work of Fraser et al. [14]. These features can be categorized as (a) syntactic, (b) readability, and (c) lexical diversity. As the name suggests, syntactic features provide information about the syntax of written communication. In our work, we used SpaCy toolkit [3] to compute normalized histogram counts of parts-of-speech and dependency tags for the transcript of each subject. The second type of handcrafted features used in this work measured the readability of subjects' transcripts. We suggest that there are differences in the readability of speech transcripts between healthy subjects and those with dementia. Hence, we used the Readability toolkit [4] to compute eight features that quantify the readability of speech transcripts. Finally, given that Alzheimer's dementia disorder affects memory, we posit that subjects with dementia will use a repetitive and less diverse vocabulary compared to healthy subjects. To quantify the diversity of their vocabulary, we computed ten features based on text-to-token-ratio using the Lexical Diversity toolkit [5].

[2]https://www.microsoft.com/en-us/microsoft-365/microsoft-stream

[3]https://spacy.io

[4]https://pypi.org/project/readability

[5]https://pypi.org/project/lexical-diversity

Table 1: *Summary of Results of Task 1 for the Training partition*

Preproc.-ID	Feature Name	Acc.	avg-f1score	Sens.	Spec.
E	facebook_bart_base__AvgPool	84.34	84.34	79.31	89.87
B	distilroberta_base__MaxPool	83.73	83.72	82.76	84.81
B	facebook_bart_base__Median	82.53	82.52	80.46	84.81
B	bert_base_multilingual_uncased__AvgPool	81.93	81.93	78.16	86.08
D	bert_base_multilingual_uncased__AvgPool	81.93	81.92	80.46	83.54
B	bert_large_uncased__MaxPool	81.93	81.89	82.76	81.01
A	facebook_bart_base__Median	81.93	81.86	83.91	79.75

Table 2: *Summary of Results of Task 2 for the Training partition*

Preproc.-ID	Feature Name	RMSE	MAE	Pearson's r
B	bert_base_multilingual_uncased__MaxPool	4.64	3.65	0.75
D	facebook_bart_base__AvgPool	4.86	3.75	0.72
A	distilbert_base_uncased__RangePool	4.87	3.84	0.72
B	distilbert_base_uncased__RangePool	4.90	3.92	0.71
D	bert_base_multilingual_uncased__RangePool	4.91	3.86	0.71

Table 3: *Summary of Results of Task 3 for the Training partition*

Preproc.-ID	Feature Name	Acc.	avg-f1score	Sens.	Spec.
C	facebook_bart_base__AvgPool	83.56	73.49	53.33	91.38
C	bert_large_uncased__MaxPool	83.56	70.08	40.00	94.83
D	facebook_bart_base__AvgPool	83.56	70.08	40.00	94.83
C	bert_base_multilingual_uncased__RangePool	84.93	69.41	33.33	98.28
C	facebook_bart_base__Median	80.82	69.07	46.67	89.66

In addition to the above, we investigated the efficacy of deep textual embeddings, such as Bidirectional Encoder Representations from Transformers (BERT) [15] and its derivatives. These models use multi-headed self-attention [16] based encoder and decoder layers which enable them to learn sophisticated latent representations from text [15, 17, 18]. Jawahar et al. [18] have shown that transformer-based models can capture structural and linguistic properties of the English language as classical tree-like structures.

We surmise that such models can represent linguistic characteristics of speech and as such be useful for differentiating between speech transcripts of healthy subjects and those with dementia. To this end, we experiment with embeddings generated using nine pre-trained transformer-based models which include: $bert\_base\_uncased$, $bert\_large\_uncased$, $distilbert\_base\_uncased$ [19], $roberta\_base$, $roberta\_large$ [20], $distilroberta\_base$,$bert\_base\_multilingual\_uncased$, $allenai\_biomed\_roberta\_base$ [21], and $facebook\_bart\_base$ [22]. We used the Huggingface library [11] in order to compute these embeddings.

It should be mentioned here that these embeddings are computed for each input token (for example, a word), and not the entire transcript as a single entity. Therefore, to generate a single feature vector for the entire transcript, we used functionals of descriptive statistics for pooling. For example, average pooling (AvgPool), maximum value pooling (MaxPool), percentile-based range pooling (RangePool), and median value pooling (MedianPool) are used in this work. The resultant feature vector is passed down to the machine learning pipeline as shown in Figure 1.

3.4. Classification and Regression

As mentioned earlier, the ADReSSo challenge 2021 consists of two classification tasks (Task 1 and Task 3) and one regression task (Task 2). We used a logistic regression classifier (LRC) and support vector machine classifier (SVC) with a linear kernel for Tasks 1 and 3, whereas for Task 2, we used partial least squares regressor (PLSR). We have previously used success using these tools [4, 10, 23]. The regularization parameter 'C' for LRC and SVC was optimized using leave-one-subject-out cross-validation (LOSO-CV) over a logarithmically spaced grid between 10^{-7} and 10^{3} whilst using the avg-f1score as the metric of classification performance. It should be mentioned that although the official performance metric for Task 1 is accuracy, we decided to use avg-f1score to optimize the regularization parameter. This was done due to the class imbalance in the training partition for the dataset provided for Task 1. Meanwhile, the 'number of components' hyper-parameter for PLSR was optimized using LOSO-CV over a grid between 2 and 40 to minimize the RMSE score. We used the scikit-learn toolkit [24] for training models for classification and regression.

4. Experiments and Results

4.1. Predictions for the training partition

In Table 1, we report the results of the top-5 models for the training partition of Task 1. Here, one can note that the best performing model uses Preprocessing-E with DTE features com-

puted from the $BART\_basemodel$. It achieves an accuracy of 84.34% with a specificity of 89.87% but a relatively poor sensitivity of 79.31%. Meanwhile, the second-placed model achieved an accuracy of 83.73% on the training partition but does not use any special processing to integrate silence information into speech transcripts. It is interesting to note that this model offers an improved sensitivity at the cost of decreased specificity when compared with the best-performing model.

In Table 2, we summarize the results for training partition of Task 2. The objective here is to predict the MMSE score assigned to subjects from the Diagnosis dataset. We report that the best performing model achieves an RMSE of 4.64 compared to the baseline of 6.42, which is a significant reduction. It is also interesting to note that there does not appear to be any advantage of integrating silence information into the speech transcript for Task 2, however, this requires further investigation.

Finally, the results for Task 3 have been summarized in Table 3 where our best performing model achieves an avg-f1score of 73.49% compared to the challenge baseline of 66.67%. Most importantly, all of the top-5 models use special preprocessing steps.

4.2. Predictions for the test partition

For Task 1, we first used confidence-based fusion to combine predictions from the top-3 performing models. This step achieved an accuracy of 81.69% for the test partition. Next, we used label-fusion of top-5 models and this increased the classification accuracy to 83.10%. Finally, we attempted label-fusion of five models selected on the basis of their specificity and sensitivity scores for the training partition. Two of these models are $facebook\_bart\_base\_\_AvgPool$ and $bert\_base\_multilingual\_uncased\_\_AvgPool$ which had provided high specificity for the training partition, whereas the remaining three, i.e. $distilroberta\_base\_\_MaxPool$, $bert\_large\_uncased\_\_MaxPool$, and $facebook\_bart\_base\_\_Median$ had provided high sensitivity. We assume that the fusion of models with high specificity only (as is the case with top-5 models) will bias predictions towards a particular class and therefore lead to poorer results overall. The resultant predictions for this fusion method achieve the best results for Task 1 where we achieved a classification accuracy of 84.51% compared to the challenge baseline of 77.46%, which is a significant improvement.

In Table 5, we summarize the results for the test partition of the dataset for Task 2. In our first attempt, we used predictions for the test partition as generated by the best-performing model for the training partition. This yielded an RMSE of 4.93 for the test partition. Our second and third attempts used averaging and median-based fusion with predictions from the top-3 models for the training partition. With these methods, we achieved an RMSE score of 4.71 and 4.54, respectively. Finally, we attempted averaging and median-based fusion for fourth and fifth attempts and achieved RMSE scores of 4.45 and 4.35, respectively. It is interesting to note that all of our attempts at predicting MMSE score achieved better performance than the challenge baseline of 5.28.

Finally, in Table 6, we summarize the results for Task 3, where the objective was to identify whether the subject has suffered from a cognitive decline over two years. Here, our best result for the test partition is an avg-f1score of 73.80%. This was achieved via predictions generated by the model which yielded the best performance for the training partition. We also experimented with label-fusion of predictions for the top-3 models for the training partition, however, this led to a decrease in avg-f1score. It should be mentioned here that we did not experiment with class-imbalance aware fusion, as used for Task 1, although in hindsight it may have been a better option.

Table 4: *Summary of Results of Task 1 for the Test partition*

Predictions (source)	**Acc.**	**avg-f1score**
baseline	77.46	–
Conf. Fusion of Top-3 models	81.69	81.64
Label Fusion of Top-5 models	83.10	82.94
Label Fusion selected models	*84.51*	84.45

Table 5: *Summary of Results of Task 2 for the Test partition*

Predictions (source)	**RMSE**
baseline	5.28
Single best model	4.93
Average-value fusion from Top-3 models	4.71
Median-value fusion from Top-3 models	4.54
Average-value fusion from Top-5 models	4.45
Median-value fusion from Top-5 models	*4.35*

Table 6: *Summary of Results of Task 3 for the Test partition*

Predictions (source)	**Acc.**	**avg-f1score**
baseline	–	66.67
Single best model	78.13	*73.80*
Label Fusion of Top-3	68.75	54.29

5. Conclusions

Alzheimer's dementia is a disease that greatly reduces the quality of life of those who suffer from it. Early detection of this disorder may assist to enhance the quality of their day-to-day lives. The purpose of the ADReSSo challenge was to develop an automated screening tool for dementia recognition. In this paper, we proposed a system based on functionals of deep textual embeddings and benchmarked its performance against the official baselines set by the ADReSSo challenge organizers. We also demonstrated that one can enrich speech transcripts with silence segments in speech recordings to yield improved performance. Overall, our proposed solution for the ADReSSo challenge offers a significant improvement for all three tasks as follows. For Task 1, we achieved an accuracy of 84.51% compared to the baseline of 77.46%, for Task 2, we achieved an RMSE of 4.35 compared to the baseline of 5.28, and for Task 3, we achieved an avg-f1score of 73.80% compared to the baseline of 66.67%. These results are a testament to the efficacy of our proposed system.

6. Acknowledgements

We would like to thank Abbas Syed for useful, constructive discussions, and assistance with running some of the experiments.

7. References

[1] World Health Organisation, "Dementia: Key Facts," 2020. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dementia

[2] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[3] Z. S. Shah, K. Sidorov, and D. Marshall, "Psychomotor Cues for Depression Screening," in *IEEE International Conference on Digital Signal Processing (DSP)*, 2017, pp. 1–5.

[4] Z. S. Syed, K. Sidorov, and D. Marshall, "Automated Screening for Bipolar Disorder from Audio/Visual Modalities," in *ACM International Workshop on Audio/Visual Emotion Challenge (AVEC)*, 2018, pp. 39–45.

[5] Z. S. Syed, S. A. Memon, and A. L. Memon, "Deep Acoustic Embeddings for Identifying Parkinsonian Speech," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, pp. 726–734, 2020.

[6] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, "Artificial Intelligence, Speech, and Language Processing Approaches to Monitoring Alzheimer's Disease: A Systematic Review." *Journal of Alzheimer's disease : JAD*, vol. 78, no. 4, pp. 1547–1574, 2020.

[7] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Detecting cognitive decline using speech only: The ADReSSo Challenge," *medRxiv*, pp. 1–5, 2021.

[8] ——, "Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge," in *INTERSPEECH*, 2020, pp. 2172–2176.

[9] B. Macwhinney, *The CHILDES project: Tools for analyzing talk*, 3rd ed. Psychology Press, 1992.

[10] M. S. S. Syed, Z. S. Syed, M. Lech, and E. Pirogova, "Automated Screening for Alzheimer's Dementia through Spontaneous Speech," in *INTERSPEECH*, 2020, pp. 2222–2226.

[11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *arXiv:1910.03771*, pp. 1–11.

[12] S. Team, "Silero Models: pre-trained enterprise-grade STT / TTS models and benchmarks," 2021. [Online]. Available: https://github.com/snakers4/silero-models

[13] J. Yuan, Y. Bian, X. Cai, J. Huang, Z. Ye, and K. Church, "Disfluencies and Fine-Tuning Pre-trained Language Models for Detection of Alzheimer's Disease," in *INTERSPEECH*, 2020, pp. 2162–2166.

[14] K. C. Fraser, J. A. Meltzer, and F. Rudzicz, "Linguistic features identify Alzheimer's disease in narrative speech," *Journal of Alzheimer's Disease*, vol. 49, no. 2, pp. 407–422, 2015.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint:1810.04805v2*, vol. 1, no. 1, pp. 1–16, 2018.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 1–11.

[17] Y. Goldberg, "Assessing BERT's syntactic abilities," *arXiv:1901.05287*, pp. 1–4, 2019.

[18] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3651–3657.

[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv:1910.01108*, pp. 1–5, 2019.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint:1907.11692*, pp. 1–13, 2019.

[21] S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," *arXiv preprint:2004.10964*, pp. 1–19, 2020.

[22] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "{BART}: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[23] M. S. S. Syed, E. Pirogova, and M. Lech, "Prediction of Public Trust in Politicians Using a Multimodal Fusion Approach," *Electronics*, vol. 10, no. 11, pp. 1–13, 2021.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.