INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Automatic Detection of Off-topic Spoken Responses Using Very Deep Convolutional Neural Networks

Xinhao Wang[1], *Su-Youn Yoon*[2], *Keelan Evanini*[2], *Klaus Zechner*[2], *Yao Qian*[1]

Educational Testing Service
[1]90 New Montgomery St. #1450, San Francisco, CA 94105, USA
[2]660 Rosedale Rd., Princeton, NJ 08541, USA
{xwang002, syoon, kevanini, kzechner, yqian}@ets.org

Abstract

Test takers in high-stakes speaking assessments may try to inflate their scores by providing a response to a question that they are more familiar with instead of the question presented in the test; such a response is referred to as an off-topic spoken response. The presence of these responses can make it difficult to accurately evaluate a test taker's speaking proficiency, and thus may reduce the validity of assessment scores. This study aims to address this problem by building an automatic system to detect off-topic spoken responses which can inform the downstream automated scoring pipeline. We propose an innovative method to interpret the comparison between a test response and the question used to elicit it as a similarity grid, and then apply very deep convolutional neural networks to determine different degrees of topic relevance. In this study, Inception networks were applied to this task, and the experimental results demonstrate the effectiveness of the proposed method. Our system achieves an F1-score of 92.8% on the class of off-topic responses, which significantly outperforms a baseline system using a range of word embedding-based similarity metrics (F1-score = 85.5%).

Index Terms: off-topic spoken responses, spoken language proficiency assessment, very deep convolutional neural networks, Inception networks

1. Introduction

This study aims to address one typical issue related to off-topic responses in the domain of spoken language assessment. In the context of large-scale, standardized assessments of spoken English for academic purposes, such as the TOEFL iBT test [1], the Pearson Test of English Academic [2], and the IELTS Academic assessment [3], some test takers may attempt to inflate their scores by modifying the topic of their responses to a topic that is more familiar to them but unrelated to the test question. These off-topic responses are thus not an authentic representation of test takers' speaking proficiency skills for the assigned topics. Therefore, in order to guarantee the validity of assessment scores, it is necessary to have a mechanism to flag these responses before scores are reported; this is especially important for an automated scoring system since automated systems tend to be more vulnerable than human raters to scoring inaccuracy due to off-topic responses [4, 5]. For example, off-topic responses may result in inflated scores from an automated system that evaluates aspects of the responses such as pronunciation, fluency, vocabulary, grammar, etc., but disregards topical irrelevance.

Researchers have examined various types of methods for the task of off-topic detection for both written essays and spoken responses. For example, [6] used vector space models (relying on exact word matching) to measure the topic relevance between written essays and test questions, and [7] furthered improved this approach by correcting spelling errors in essays and expanding words in the test questions with inflected forms, synonyms, and distributionally similar words.

In the domain of off-topic detection for spoken responses, [8] detected non-scorable responses based on vector space models, where hundreds of high-proficient responses were collected according to each question and used as reference samples for model building. Instead of using recognized words for the off-topic detection of spoken responses, [9] built a model for the Pearson Test of English Academic by using features derived from speech confidence scores. This approach achieved good performance for restricted speech, but it is not appropriate for tasks that elicit unconstrained speech.

More recently, deep neural networks and word embeddings have become the dominant approaches for measuring semantic similarity between two sentences/documents [10, 11]. [12] provided sentence-level relevance scores for written essays by using various similarity metrics based on word embeddings. [13] adapted a Recurrent Neural Network language model to the topic of each question with sample responses and then ranked the topic-conditional posterior probabilities of a spoken response. [14] investigated three types of similarity metrics based on word embeddings as well as Siamese Convolution Neural Network for off-topic detection of spoken responses and showed the effectiveness of the proposed similarity metrics on this task. In view of the work in [14], we built a baseline system with three types of similarities based on word embeddings and compared it with the proposed method.

In general, models trained with reference samples from each test question can achieve the best results. However, it is not always feasible to collect such samples in advance, for example, when new test questions are launched. Therefore, in consideration of the demands of operational deployment, this study aims to build an automatic off-topic detection model, where no reference samples are used and only test question prompts are available for model building. We propose to construct a similarity grid to measure the semantic similarities between the word sequence in a test response and that in the assigned question prompt. Afterward, inspired by the recent progress of technologies on the challenging task of image recognition, state-of-the-art very deep neural networks were employed to differentiate off-topic responses from on-topic ones.

Recurrent neural networks (RNN) and convolutional neural networks (CNN) have been widely used for many tasks in the fields of speech and natural language processing (NLP). However, compared with the very deep convolutional networks

http://dx.doi.org/10.21437/Interspeech.2019-1848

widely used in the field of computer vision, these architectures are rather shallow. [15] proposed an analogy that the hierarchical structure of texts (from characters to stems, words, phrases, sentences, etc.) is similar to the compositional structure of images (hierarchically assembling pixels into objects). Based on this property, very deep CNNs could potentially benefit NLP tasks as they have for image recognition. [15] used up to 29 convolutional layers in their network to perform text classification at the character level and achieved improvements over the state-of-the-art on several public corpora. This study, however, directly represents the comparison between two documents in a visual similarity grid and then employs state-of-the-art techniques for image recognition to determine their similarity.

2. Similarity Grid

This study aims to measure the topical relevance of a response solely by comparing the response to the text of the test question that was used to elicit the test taker's response; this is referred to as the prompt text hereafter. In addition, a preprocessing step was required to remove all stop words from both responses and prompts, and thus the semantic comparison was conducted only on the remaining content words. We propose to construct a similarity grid for each pair of a response and the corresponding prompt, in which the content word sequence from the response is included on the y-axis from top to bottom, and the content word sequence from the prompt is included on the x-axis from left to right. Accordingly, a cell (i, j) in the grid indicates a similarity measurement between the i^{th} content word in the response and the j^{th} content word in the prompt. In this work, the semantic similarities of word pairs are calculated as the cosine similarity between word embeddings, where the word2vec model [16] trained on the Google News Corpus[1] was used to extract embedding vectors.

The similarity grids have one channel, i.e., one single measurement value for each cell, and they can be visualized as grayscale images with lighter cells (pixel values closer to 255) indicating higher degrees of similarity and darker cells (pixel values closer to 0) indicating lower similarity. Figure 1 shows similarity grids for two example responses to the same test question: the image on the left corresponds to an on-topic response and image on the right corresponds to an off-topic response. The comparison between these two images indicates that more cells in the on-topic grid are lighter than in the off-topic one; accordingly, we can interpret the task of off-topic detection as filtering grids with more darkness and less brightness.

Furthermore, just as in composing an image, the similarity grid can be represented in grayscale with one channel (each pixel in the image is encoded with only one value) or with multiple channels, as in an RGB image with 3 channels (each pixel is encoded with three different values, one value corresponding to each channel). Therefore, additional channels can be used in the similarity grid to convey additional information comparing between the response and the prompt; for example, in addition to semantic similarity values, other metrics measuring word importance can be stored in other channels. Inverse document frequency, idf, weights have been widely used to indicate the importance of different words in a document in tasks such as text classification and information retrial. Here, based on idf values, 1-channel grids can be expanded to 3-channel ones. For example, for each cell (i, j) in a grid, the value in the first channel can still be the cosine similarity of word embeddings; the value in the second channel can be the idf weight of the i^{th} word in the response; similarly, the value in the third channel can be the idf weight of the j^{th} word in the prompt. In this way, the similarity measurement at each cell can be scaled in terms of idf word importance values. In fact, the number of channels is not limited, and a similarity gird can consist of as many channels as necessary to encode the inputs from different aspects.

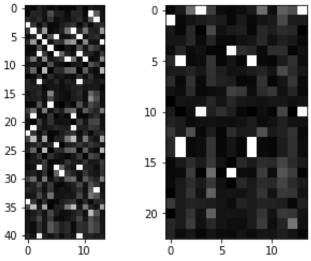

Figure 1: *Visualization of two sample similarity grids. The left and right figures correspond to on-topic and off-topic responses to the same test question, respectively.*

Due to the large variations in the lengths of test responses and prompt texts, the sizes of the similarity grids fluctuate substantially. In order to meet the constraint of fixed-length input for the Inception networks, a commonly used image resizing method based on bilinear interpolation was applied to scale all similarity grids into a standard size of 180 (the maximum length of a test response) by 180 (the maximum length of a prompt).

3. Inception Networks

In the past several years, researchers have achieved huge advances in the field of computer vision by introducing very deep convolutional neural networks, for example, AlexNet [17], VGG16 [18], GoogLeNet (Inception-v1) [19], BN-Inception-v2 [20], Inception-v3 [21], Inception-v4 [22], Inception-ResNet [22], as well as Squeeze-and-Excitation networks [23]. Each time, these successive models continue to show improvements by validating their work against the ImageNet dataset[2] and Challenges. In particular, the evolution of Inception networks was an important milestone that enhanced the performance in terms of both accuracy and speed. Besides obtaining state-of-art performance on various image classification tasks in the computer vision community, Inception networks have also been introduced into other fields. For example, Zhang et al. [24] used the Inception-Resnet-v1 [22] (an Inception network with residual connections) to extract speaker embeddings for the task of text-independent speaker verification.

The Inception network consists of a highly hand-crafted architecture; Figure 2 shows the schema of one example Inception network, the Inception-v4. "Stem" is an initial set of stacked convolution/max-pooling operations performed before applying Inception blocks, and it can vary across different versions of Inception networks. In Inception-v4, there are three main modules as well as one reduction block. Figure 2 also shows an example of the Inception-A module.

The main characteristics of the Inception modules are as follows. First, the kernel size of convolution operations relates to the range of distributed information that is captured by filters,

[1] Downloaded from `https://code.google.com/archive/p/word2vec/`

[2] `http://www.image-net.org/`

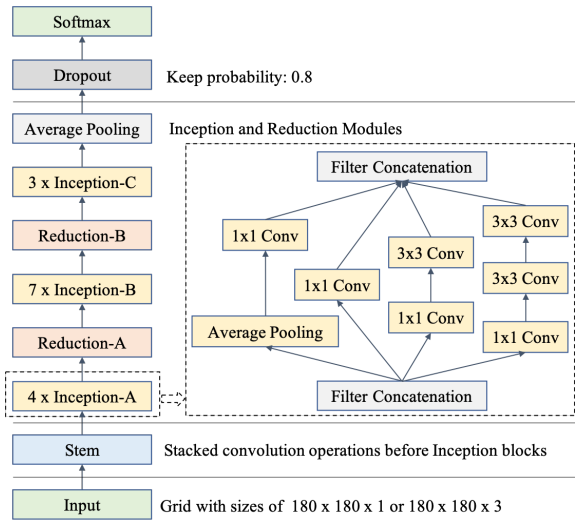

Figure 2: *The schema of the Inception-v4 network. One example Inception module (Inception-A) is also displayed. Detailed implementation for each block can be found in [22].*

i.e., the larger more globally and the smaller more locally. Inception modules avoid the tough decision of choosing the right size of kernels by having multiple different sizes of filters in parallel at the same level. Thus Inception networks are wider in addition to being deeper. Second, in order to reduce the expensive computation, a challenge that is always faced by very deep neural networks, Inception modules reduce the dimension of input channels by adding an extra 1×1 convolution before larger convolutions. In addition, they also use factorization to break down convolutions with larger sizes into smaller ones, for example, factorizing a 5×5 convolution into two consecutive 3×3 ones; factorizing a $n \times n$ convolution into two consecutive ones with sizes of $1 \times n$ and $n \times 1$, respectively. Furthermore, inspired by the work proposed by He et al. [25], residual connections are introduced into Inception-ResNet, which can speed up the training process for very deep networks.

There exist many other operations introduced during the evolution of Inception networks, and more implementation details can be found in the literature [19, 20, 21, 22]. In this work, we investigated three versions for the task of off-topic detection: Inception-v3, Inception-v4, and Inception-ResNet-v2. The TensorFlow source code[3] was used to develop models.

4. Experiments

This work focuses on the task of off-topic response detection in the domain of a large-scale, high-stakes assessment of English for non-native speakers which assesses English communication skills for academic purposes. The Speaking section of this assessment contains six tasks designed to elicit spontaneous spoken responses. Two of the tasks use questions that ask about test takers' information or opinions on familiar topics based on their personal experience or background knowledge; these are referred to as independent tasks. The other four tasks require test takers to summarize or discuss material provided in a reading and/or listening passage; these are referred to as integrated tasks. In general, the independent tasks ask questions on topics that are familiar to test takers and are not based on any stimulus materials. A sample independent question is "Talk about an activity you enjoyed doing with your family when you were a kid". Therefore, test takers can provide responses containing a wide variety of specific examples, and most instances of off-topic responses were found in response to these independent questions.

[3] https://github.com/tensorflow/models/tree/master/research/slim

Table 1: *Distribution of prompt texts in terms of number of words and number of content words (after removing stop words)*

	min	max	mean	std
number of words	9	60	31.2	10.0
number of content words	4	35	15.7	5.4

4.1. Experimental Setup

In order to conduct this study, we collected a large number of spoken responses from operational administrations of the assessment. All of them were elicited using independent questions and each response contained approximately 45 seconds of spontaneous speech from non-native speakers of English. A total of 283 questions covering a wide range of topics such as education, entertainment, health, and policies were used in this study. The prompt texts presented to test takers in these questions were relatively short and typically consisted of just a few sentences. Table 1 shows that the number of words in each prompt text ranges from 9 to 60. After removing stop words, the shortest prompt text includes only 4 content words.

We collected $183,111$ spoken responses elicited with the 283 questions described above and further partitioned them into two sets: 120,115 in the Training set and 62,996 responses in the Test set. There was no speaker overlap between the two partitions. All responses used in this study were originally scored by expert human raters during the operational test, and off-topic responses are rare in such a scoring scenario. Since it is not very practical to collect a large amount of authentic off-topic responses from actual administrations of the test, we created a set of synthetic off-topic responses for the following experiments. Each question in our assessment was designed to elicit content that was substantially different from others, and therefore, mismatched responses have substantial content issues, i.e. a response to one question is not topically related to another question. Furthermore, experts (assessment developers) suggested that test takers could recite pre-memorized responses (for different questions) regardless of which question they were given. According to this assumption, within each test question, we randomly selected a subset from responses elicited with the other 282 questions and took them as off-topic responses for this give question. Among each partition, the same number of off-topic responses were selected according to the number of on-topic responses, resulting in a ratio of 1:1 between on-topic and off-topic responses.

A Kaldi[4]-based automatic speech recognition (ASR) engine [26], which had a word error rate (WER) of around 23% on a held-out test set with 600 responses, was employed to transcribe the non-native speech into text. The ASR system consisted of a gender-independent acoustic model and a trigram language model, which were trained with a data set including similar responses (around 800 hours of speech) drawn from the same assessment.

[4] http://kaldi-asr.org/

4.2. Baseline System

Following the previous work in [14], which demonstrated that similarity features based on word embeddings can outperform a Siamese CNN, we built a baseline system with the following three different types of features:

- Word Mover's Distance (WMD): This feature calculates the sum of the minimum distances between words in the two compared documents (a test response and a prompt text) where the distance between two words was the Euclidean distance between the two corresponding word vectors in the embedding space [10].
- Averaged word embeddings: Given an input document, a representative vector can be generated by mapping each word to its embedding vector and then averaging all word vectors. Then the cosine similarity between two vectors representing a test response and a prompt text can be calculated.
- *idf*-weighted word embeddings: When generating the representative vector for an input document, *idf* weights can be used to scale each word embedding; then, the weighted embeddings can be averaged and the cosine similarity between a test response and a prompt text can be calculated.

These features measured the semantic similarity between a response and a test question in an embedding space, where the *word2vec* model used in constructing similarity grids in Section 2 was also used to extract word embeddings, and the *gensim* package [27] was used to calculate the WMD. Finally, the baseline system was built with a Random Forest classifier[5] using the *scikit-learn* machine learning toolkit [28].

4.3. Results and Discussion

The proposed method based on similarity grids and Inception networks was compared with the baseline system. As shown in Table 2, the baseline system obtained the lowest F1-score of 85.5%. When constructing the similarity grids without *idf* values, Inception-v4 can achieve the best F1-score at 89.1%. Furthermore, by appending *idf* channels into grids, the F1-scores can be consistently improved across all three Inception networks, and Inception-Resnet-v2 achieves the best F1-score at 92.8%, substantially outperforming the baseline system. With *idf* weights to indicate word importance in the similarity grid, the precision of Inception-Resnet-v2 was markedly increased from 85.6% to 91.5%, along with a 3.1% improvement on the recall. Also as reported in [22], the addition of residual connections into Inception-Resnet-v2 can speed up the training process by making it converge with fewer epochs.

Furthermore, we break down the F1-scores according to the lengths of prompt texts (number of content words included in the test questions). As shown in Figure 3, despite certain fluctuations, automatic systems tend to perform better on questions with more content words, and Inception-Resnet-v2 consistently outperforms the baseline across all prompts. In particular, with only very limited numbers of content words in the prompts, for example, less than 10, the Inception network can achieve larger gains. Meanwhile, with the longest test question (including 35 content words), the improvement with Inception-Resnet-v2 is also comparably larger.

[5]Random Forest Classifier was selected because of its superior performance over different machine learning algorithms in a pilot experiment

Table 2: *Precision, recall, and F1-score on the off-topic class with different models.*

	Precision (%)	Recall (%)	F1 (%)
Similarity Grid without *idf*			
Inception-v3	84.7	92.2	88.3
Inception-v4	**86.5**	**91.8**	**89.1**
Inception-Resnet-v2	85.6	91.1	88.3
Similarity Grid with *idf*			
Inception-v3	90.2	94.1	92.1
Inception-v4	90.9	93.4	92.1
Inception-Resnet-v2	**91.5**	**94.2**	**92.8**
Baseline	81.4	90.0	85.5

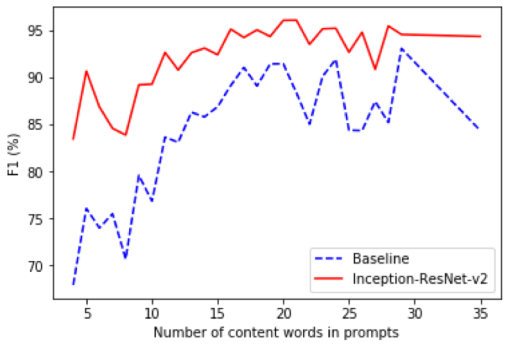

Figure 3: *F1-scores compared to the length of prompt texts.*

5. Conclusion and Future Work

This paper proposes an innovative method to detect off-topic responses in the context of spoken language assessment. Similarity grids were constructed to capture the topical relatedness between a test response and a test question, and then very deep convolutional neural networks, i.e., three versions of Inception networks, were applied to build detection models. Compared with the baseline system, the proposed method can increase the absolute F1-score on the off-topic class by 7.2%. Although the prompts from the test questions used in our work are very short, and only a limited number of content words can be used for the similarity measurement, we can still achieve a promising F1-score of 92.8%.

In fact, the approach presented in this study can be applied to any task that relies on similarity measurements between two documents. In particular, there exists another type of response that can also impact the validity of spoken language assessment, i.e., plagiarized spoken responses. Some test takers may attempt to game the test by memorizing prepared source materials before the test and then adapting them on-the-fly during the test to produce their spoken responses. These responses can be either on-topic or off-topic. When trying to identify such instances of plagiarism, experienced human raters attempt to find salient matching expressions that appear both in potential source materials and the test responses. In our future work, we also plan to visualize a grid of exact lexical matches between a test response and a source and then build very deep convolutional neural networks to detect such salient matching expressions.

6. References

[1] ETS, *The Official Guide to the TOEFL® Test, Fourth Edition*. McGraw-Hill, 2012.

[2] P. Longman, *The Official Guide to Pearson Test of English Academic*. Pearson Education ESL, 2010.

[3] P. Cullen, A. French, and V. Jakeman, *The Official Cambridge Guide to IELTS*. Cambridge University Press, 2014.

[4] K. E. Lochbaum, M. Rosenstein, P. Foltz, and M. A. Derr, "Detection of gaming in automated scoring of essays with the IEA," *Presented at 75th Annual meeting of NCME*, 2013.

[5] D. Higgins and M. Heilman, "Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior," *Educational Measurement: Issues and Practice*, vol. 33, no. 3, pp. 36–46, 2014.

[6] D. Higgins, J. Burstein, and Y. Attali, "Identifying off-topic student essays without topic-specific training data," *Natural Language Engineering*, vol. 12, no. 2, pp. 145–159, 2006.

[7] A. Louis and D. Higgins, "Off-topic essay detection using short prompt texts," in *Proceedings of the NAACL-HLT workshop on innovative use of NLP for building educational applications*, 2010, pp. 92–95.

[8] S.-Y. Yoon and S. Xie, "Similarity-based non-scorable response detection for automated speech scoring," in *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, 2014, pp. 116–123.

[9] J. Cheng and J. Shen, "Off-topic detection in automated speech assessment applications," in *Proceedings of Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[10] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *Proceedings of International Conference on Machine Learning*, 2015, pp. 957–966.

[11] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proceedings of Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[12] M. Rei and R. Cummins, "Sentence similarity measures for fine-grained estimation of topical relevance in learner essays," in *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, 2016.

[13] A. Malinin, R. C. Van Dalen, Y. Wang, K. M. Knill, and M. J. Gales, "Off-topic response detection for spontaneous spoken english assessment."

[14] S.-Y. Yoon, C. M. Lee, I. Choi, X. Wang, M. Mulholland, and K. Evanini, "Off-topic spoken response detection with word embeddings," in *Proceedings of INTERSPEECH*, 2017, pp. 2754–2758.

[15] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," in *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 2017.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of Advances in neural information processing systems*, 2013, pp. 3111–3119.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., 2012, pp. 1097–1105.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, 2014.

[19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, 2014.

[20] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, 2015.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *CoRR*, 2015.

[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[24] C. Zhang, K. Koishida, and J. H. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 9, pp. 1633–1644, 2018.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[26] J. Tao, S. Ghaffarzadegan, L. Chen, and K. Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *Proceedings of ICASSP*. IEEE, 2016, pp. 6140–6144.

[27] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, pp. 2825–2830, 2011.