Towards Scalable Vocabulary Acquisition Assessment with BERT

Zhongdi Wu
zhongdiw@smu.edu
Southern Methodist University
Dallas, Texas, United States

Eric Larson
eclarson@lyle.smu.edu
Southern Methodist University
Dallas, Texas, United States

Makoto Sano
Makoto.Sano@mack-the-psych.com
Mack-the-Psych.com
Setagaya, Tokyo, Japan

Doris Baker
doris.baker@austin.utexas.edu
University of Texas at Austin
Austin, Texas, United States

Nathan Gage
ngage@mail.smu.edu
Southern Methodist University
Dallas, Texas, United States

Akihito Kamata
akamata@smu.edu
Southern Methodist University
Dallas, Texas, United States

ABSTRACT

In this investigation we propose new machine learning methods for automated scoring models that predict the vocabulary acquisition in science and social studies of second grade English language learners, based upon free-form spoken responses. We evaluate performance on an existing dataset and use transfer learning from a large pre-trained language model, reporting the influence of various objective function designs and the input-convex network design. In particular, we find that combining objective functions with varying properties, such as distance among scores, greatly improves the model reliability compared to human raters. Our models extend the current state of the art performance for assessing word definition tasks and sentence usage tasks in science and social studies, achieving excellent quadratic weighted kappa scores compared with human raters. However, human-human agreement still surpasses model-human agreement, leaving room for future improvement. Even so, our work highlights the scalability of automated vocabulary assessment of free-form spoken language tasks in early grades.

CCS CONCEPTS

• **Applied computing** → **Computer-managed instruction**; • **Computing methodologies** → **Neural networks**.

KEYWORDS

Automated Scoring, Human-Machine Reliability, Deep Neural Networks, Transfer Learning, Natural Language Processing

ACM Reference Format:
Zhongdi Wu, Eric Larson, Makoto Sano, Doris Baker, Nathan Gage, and Akihito Kamata. 2023. Towards Scalable Vocabulary Acquisition Assessment with BERT. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23), July 20–22, 2023, Copenhagen, Denmark.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3573051.3596170

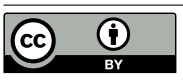
1 INTRODUCTION

Understanding how well English language learners understand academic vocabulary while they develop language proficiency is an important goal for educators and researchers. One method of assessment that is theoretically and empirically based, and that has the potential to increase our understanding of how second language learners develop their academic vocabulary is to allow learners to respond verbally when describing or using academic terms [9]. However, scoring student speech can be very time consuming. To mitigate this problem, previous work has established an end-to-end pipeline for training and evaluating automated scoring (AS) models with hand transcribed data and word embedding based recurrent network models [15]. In parallel, transfer learning has been employed with pre-trained language models to specific tasks, achieving great success in numerous classification tasks. In most studies, transfer learning is based on pre-trained language models where a few fully connected layers are trained for a specific task, but the language model weights remain frozen (i.e., untrained) [14]. This evidence motivates us to use transfer learning of a language model for assessment of vocabulary acquisition. In particular, we investigate two separate classification tasks: one task is referred to as a "definition comprehension task" where a student is asked to define an academic word; and another task is referred to as a "sentence comprehension task" where a student is asked to use the word in a sentence. In each task, students respond verbally and transcriptions are used to assess understanding.

Thus, this study extends previous research on AS models by introducing BERT [7] pre-trained language model with transfer learning, emphasizing the importance of (1) exploring appropriate methods for supplying input vectors to pre-trained models, (2) ensuring convexity of transfer learning architectures by using skip connections, and (3) carefully selecting loss functions to better model the classification task as related to human subjective annotations. Specifically, in this study we address the following research questions:

(1) What information should be exposed to a pre-trained language model for classifying definitions and sentence usage?
(2) Which architectural elements are most influential for increasing human-machine agreement?
(3) How well does the best model (as judged by validation set performance) generalize to a test set that is constructed with words that do not appear in the training set?

Moreover, our study indicates a series of implications and potential improvements in terms of data collection, human rating process, and network architectures. These implications are currently being employed in a new round of data collection. Once this round is completed, more methods will continue to be investigated.



L@S '23, July 20–22, 2023, Copenhagen, Denmark

ACM ISBN 979-8-4007-0025-5/23/07.
https://doi.org/10.1145/3573051.3596170

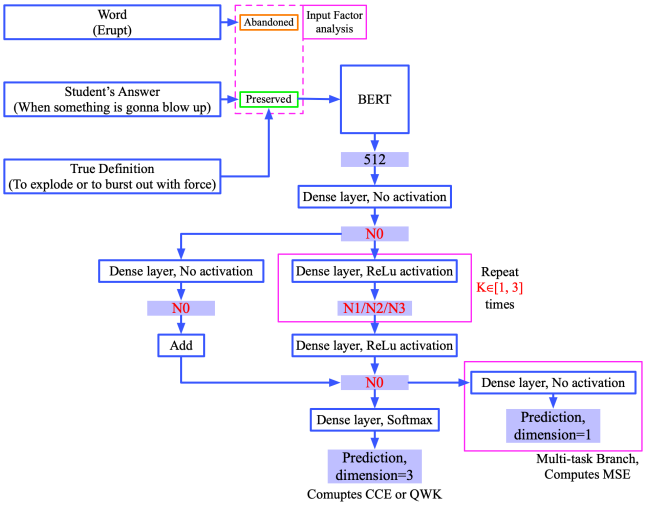


Figure 1: Example of the end-to-end pipeline of the network. Variables marked in red are found through hyperparameter search.

2 DATASET AND METHODS

We use the ELVA dataset to inform the design and evaluate our AS models [15]. Student responses in the dataset are assessed by a metric referred to as depth of knowledge (DOK) [3]. As mentioned, the measure contains two separate tasks: a definition comprehension task and a sentence comprehension task, hereafter referred to as *definition* and *sentence* tasks. In the original data collection, human annotators evaluated DOK for the definition task by scoring student responses between 0-2 and the sentence tasks from 0-3. Further analysis revealed that the inter-rater reliability (IRR) of the sentence task was greatly improved by only allowing scores from 0-2. Thus, we decided to adopt this scoring method for all responses (i.e., definitions and sentences).

A total of 13,471 English recorded utterances from 217 second grade participants were used in the study (6,778 for definition task and 6,693 for sentence comprehension). However, a substantial portion of the data was marked by annotators as "Don't Know" (DK) and "No Response" (NR), for when the student responded saying they did not know the word or when the student did not verbally respond, respectively. In our analysis we removed DK and NR responses from the dataset because they artificially boost performance evaluation. That is, these responses are relatively easy to identify by humans and the AS models, which can boost agreement measures even though the AS model is a simple pattern matching similar phrases to "I don't know." Excluding DK and NR leaves a dataset with 2,163 and 2,298 examples respectively.

2.1 Performance Measure

To evaluate the AS models, we measured the IRR between human raters and AS models by calculating the quadratic weighted kappa (QWK), as follows:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \quad \text{where} \quad w_{i,j} = \frac{(i-j)^k}{(N-1)^k} \tag{1}$$

O, w and E are the matrices of observed scores, penalty weights, and expected scores. $O_{i,j}$ corresponds to the number of responses that receive a score i from one rater and score j from another rater. Matrix E is calculated via the outer product between the score vectors of the two raters, normalized to have the same sum as Matrix O. In calculating the weight entries, $w_{i,j}$: N is the number of possible ratings (in this study $N = 3$), k indicates the strength of the penalty (since we employ quadratic kappa, $k = 2$). $\kappa = 1$ indicates perfect agreement and $\kappa = 0$ indicates no agreement.

2.2 Transfer Learning

Transfer learning refers to the methodology that improves the performance of target task model by transferring the knowledge from similar tasks (*i.e.*, classification task) within a given domain (*i.e.*, processing speech) [18]. Transfer learning has achieved great success in many domains and applications [5, 10], especially in the area of Natural Language Processing (NLP). BERT (Bidirectional Encoder Representations from Transformers) is a large language model designed for predicting the next sentence given a prompt. The encoder of BERT is often used to analyze a text sequence and provide a latent representation that is useful for many text classification tasks such as clinical text classification[8], academic paper classification[4] and fake news detection[17]. In our study, we employ BERT to analyze the transcription of student responses using several competing pre-processing methods. After BERT has generated the latent representation of the responses, fully connected neural network layers can be trained for a new classification task—in our study these layers are trained to predict the human scoring of the response, 0-2.

2.3 Loss Function and Objective Architecture

Many neural networks, including AS models in related works [15] are trained using Categorical Cross Entropy (CCE) as a loss function, l_{CCE}. CCE is often a go-to option for classification tasks. However, in this particular task, different assessment scores (0, 1, 2), though discrete, have quantitative relations. That is, the distance between score 0 and 1 is smaller than that between 0 and 2, and we would like the model to reflect this while training. Such quantitative relations between different scores are not represented in the CCE loss function. Thus, in our study we use several loss functions that account for this, such as QWK and mean squared error, in experiments and compare with CCE. We also employ multi-task learning, which allows for incorporation of multiple loss functions simultaneously.

In this study multi-task learning is implemented so that the majority of the transfer learning architecture is optimized with two loss functions. However, at the output layer, the network branches into two fully connected layers with output dimensions of one and three (*i.e.*, the number of classes) as shown in Figure 1. The single dimension branch directly predicts the target label (0-2) using the Mean Squared Error (MSE) loss, $l_{MSE} = (y - \hat{y})^2$. On the other hand, the three dimensional output is followed by a softmax and CCE is used as a loss function. The combined loss function is:

$$l_{multi-task} = \omega_{MSE} \cdot l_{MSE} + \omega_{CCE} \cdot l_{CCE} \tag{2}$$

Where ω_{MSE} and ω_{CCE} are weight multiplier hyper parameters. All hyper parameters are found using a search algorithm, as explained later. We adapt the QWK from equation 1 for use as a loss function. We first use logarithms to decouple the numerator and the denominator, which simplifies the computation of gradient [6]. Thus, the problem is reformulated into a minimization of l_{QWK}:

$$l_{QWK} = \log(1 - \kappa + \epsilon) \tag{3}$$

where κ is from equation 1 and ϵ avoids calculating $\log(0)$. The incorporation of each loss function is investigated as a hyper parameter.

2.4 Architecture and Variations

We investigate preprocessing variations in how different language prompts are provided to the BERT model. Available input prompts to BERT include (1) the word that is being asked, (2) the student's answer, and (3) the correct definition (for definition task) or example sentences (for sentence task). Intuitively not all components may needed as inputs to pre-trained language model since some potentially bring noise or convey redundant information. However, the student's answer should always be a part of the input. Thus, we use an input factor analysis (Figure 2) to select combinations that can be provided to BERT. The combinations include: (1) only using the student answer, (2) using the student answer + true definition/sentence, (3) using the student answer + word, and (4) using all inputs as prompts. For a particular combination, components were tokenized and segmented by BERT's symbol segmentation engine. The outputs of BERT for each sample in training set were then collected as the high dimensional vector representation. We investigate the quality of each combination in the next section.

Finally, as shown in Figure 1, we also used residual connections in our layers after the pre-trained BERT model. Residual connection architectures also allow the output to maintain a convex function of the output of BERT [2]. This convexity might help to increase performance of the model. The usage of residual connections is also a hyper parameter.

3 ANALYSIS AND EXPERIMENTS

Given the variations of network architectures and loss functions we introduced in this study, we systematically investigate how each aspect influences performance. First, we use visualizations to understand the input factor analysis and, second, we describe the implementation of our hyper parameter search.

3.1 Selection of input factors

Figure 2 shows visualizations of the embeddings for each competing combination of input factors. For each combination of factors, the vector representations were visualized with dimensionality reduction via stochastic neighbor embedding (t-SNE) [16] and color coded by human rating, 0-2. Within-class and between-class variance were also calculated with these vector representations (without dimensionality reduction) aiming to find a combination that has the smallest within class variance and largest between class variance (*i.e.*, representations that are most separable on the training set). With this analysis, we were able to answer the first research question:

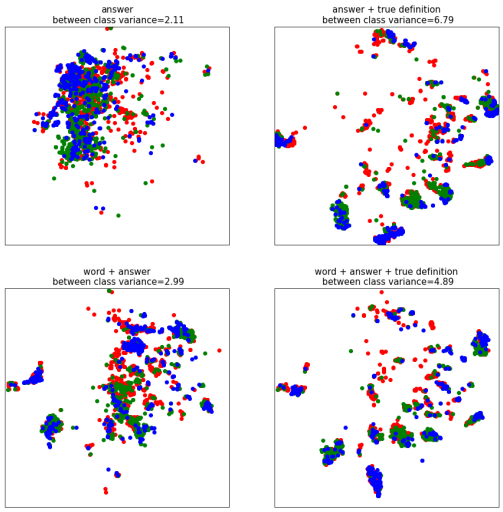


Figure 2: BERT embedding visualization with t-SNE. All subplots have the same range of x-axis and y-axis. Between class variance is noted in subplot title, as value $\times 10^{-3}$.

What information should be exposed to a pre-trained language model for classifying definitions and sentence usage? The answer is: Student's answer and the correct definition/sentence has the best separability both visually and using the within-class and between-class variance. Thus, for the remainder of our hyper parameter tuning we employ this combination only.

3.2 Other Experimental Details

Hyper-parameter searches were conducted with same train validation split on combinations of dense network/skip-connection and different objective architectures. 10-fold cross validation was carried out with the best parameters searched. The models are trained and compared without DK and NR examples in the dataset. Adaptive momentum (AdaM) optimization is employed for all training [12]. Investigations were also conducted with RMSprop and SGD, with Adam having superior performance and higher reliability in all testing. We also employ decoupled weight decay regularization to help prevent overfitting [13]. The AdaM-learning rate, number of layers, number of hidden nodes for each layer, weight decay rate, and multi-task loss weights were registered as hyper-parameters. We employ both a train-validation split of $0.8 : 0.2$ and a 10-fold cross validation split. Hyper parameter tuning was carried out using efficient sampling and pruning mechanisms with optuna [1]. All model variations were set to 2000 iterations of hyper parameter search with all hidden layers size ranging from 16 to 256. The parameters found through this search are shown in Figure 1. Searching processes were pruned if no better parameters were found in 200 iterations.

4 RESULTS

The performance of all models in terms of QWK is shown in the Table 1. The results reported are the best for each model among all hyper parameters searched. With the best performance in each

Table 1: Performance of different models using QWK

	Included Elements					Definition		Sentence	
AS model	Multi-task	Residual	QWK	CCE	MSE	80:20 Valid.	10-Fold	80:20 Valid.	10-Fold
Previous Best	-	-	-	✓	-	0.650	NULL	0.580	NULL
CCE-dense	-	-	-	✓	-	0.749	0.703 ± 0.068	0.677*	0.626 ± 0.061
QWK-dense	-	-	✓	-	-	0.771	0.751 ± 0.053	0.651	0.692 ± 0.045*
Multi-dense	✓	-	-	✓	✓	0.774	0.723 ± 0.065	0.670	0.647 ± 0.049
CCE-res	-	✓	-	✓	-	0.750	0.697 ± 0.057	0.671	0.642 ± 0.055
QWK-res	-	✓	✓	-	-	0.775	0.755 ± 0.058*	0.677*	0.665 ± 0.043
Multi-res	✓	✓	-	✓	✓	0.779*	0.737 ± 0.067	0.669	0.648 ± 0.051

column marked in Table 1, we are able to partially address our second research question: **Which architectural elements are most influential for increasing human-machine agreement?** There is not a clear overall winner, however, it is clear that the objective architectures that encode the quantitative relations (*i.e.*, using QWK) between different scores have superior performance with similar network architecture. Meanwhile, residual-connections tend to improve performance further.

4.1 Generalization Analysis

In our remaining analysis, we use the best overall models for the definition task and sentence task, which are *Multi-Res* and *QWK-Dense*, respectively. We turn our attention to the final research question: **How well does the best model (as judged by validation set performance) generalize to a test set that is constructed with words that do not appear in the training set?** We investigate this by taking out the words "erupt" and "scientist" from the training set, and re-training the models without these words. We also investigate models with half of the left-out words used in the training set in order to understand how a smaller quantity of word examples influences performance. We then evaluate on each model using the left out words.

Table 2: Performance of generalization on unseen words

	remove **all/half** "erupt" and "scientists" from train set	
Agreement between	Definition	Sentence
Rater1 and Rater2	0.96	0.84
Rater1 and Rater3	0.93	0.79
Rater2 and Rater3	0.90	0.87
Rater1 and Machine	0.71 / 0.75	0.50 / 0.68
Rater2 and Machine	0.69 / 0.75	0.53 / 0.76
Rater3 and Machine	0.69 / 0.71	0.55 / 0.72

Table 2 shows the QWK agreement among pairs of raters from the ELVA dataset. To investigate agreement, we report the QWK among pairs of raters and treat the model like a fourth rater. We observe that:(1) Including example words in training always enhances performance; (2) Human-Machine agreements are much higer on definition tasks compared to sentence tasks, models generalize better on new words for definition tasks; (3) Human-Machine agreements improvement is more apparent on sentence tasks compared to definition tasks when some data involving the words are in the training set. Therefore, sentence tasks may require vastly more examples to perform close to human level.

5 IMPLICATIONS OF FINDINGS

In our analyses, we seek implications to help guide future data collection related to vocabulary acquisition for English language learners. These implications are currently being integrated into the next round of data collection.

Problem: DK and NR with manual transcription creates many zero variation data that artificially boosts agreement. As noted by [15], manual transcription is inefficient. **Improvement**: Transcription is now automatic using cloud services. NR and DK data is automatically detected. **Problem**: Simply providing a score of 0-2 sometimes make raters not confident with their scoring. **Improvement**: Raters now have a list of rubrics to select for every score. This rubric characterizes the score so that raters can select reasoning for a certain score. Future AS models should be able to incorporate this rubric as another task during optimization. **Problem**: Words can have ambiguous meanings and usages. **Improvement**: More than one true definition could be rated against during scoring. When raters score the response, raters are asked to mark the true definition that the students are trying to lead to. **Problem**: Sentence tasks are generally harder even for human raters. The volume of valid variations is extensive. Example sentences from high scoring students can be dramatically different for a given word. **Improvement**: The sentence task now provides an image related to the word, students are asked to create a sentence related to the image.

We conclude that the limit of performance with AS modeling on the current dataset has likely been achieved. With the second round of data collection, the volume, reliability, and granularity of the data will improve. Thus, additional analysis will be conducted to better understand the problem and creating increasingly reliable models. Many investigations will be enabled, including: generalization analysis with a larger number of words; additional analysis of pre-trained language models; supervised contrastive learning [11] can be adopted; latent space vector behavior for different models can be investigated, among many others.

ACKNOWLEDGMENTS

This research was supported, by grants from the Institute of Education Sciences (R305A140471 and R305A200521). The views expressed within this article are those of the authors and do not represent the views of the United States Department of Education or the Institute of Education Sciences.

REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-Generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2623–2631. https://doi.org/10.1145/3292500.3330701

[2] Brandon Amos, Lei Xu, and J. Zico Kolter. 2017. Input Convex Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 146–155. https://proceedings.mlr.press/v70/amos17b.html

[3] Scott Baker, Lana Santoro, David Chard, Hank Fien, Yonghan Park, and Janet Otterstedt. 2013. An Evaluation of an Explicit Read Aloud Intervention Taught in Whole-Classroom Formats In First Grade. *The Elementary School Journal* 113 (03 2013), 331–358. https://doi.org/10.1086/668503

[4] Linkun Cai, Yu Song, Tao Liu, and Kunli Zhang. 2020. A Hybrid BERT Model That Incorporates Label Semantics via Adjustive Attention for Multi-Label Text Classification. *IEEE Access* 8 (2020), 152183–152192. https://doi.org/10.1109/ACCESS.2020.3017382

[5] Hang Chang, Ju Han, Cheng Zhong, Antoine M. Snijders, and Jian-Hua Mao. 2018. Unsupervised Transfer Learning via Multi-Scale Convolutional Sparse Coding for Biomedical Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 5 (2018), 1182–1194. https://doi.org/10.1109/TPAMI.2017.2656884

[6] Jordi de la Torre, Domenec Puig, and Aida Valls. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters* 105 (2018), 144–154. https://doi.org/10.1016/j.patrec.2017.05.018 Machine Learning and Applications in Artificial Intelligence.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[8] Shang Gao, Mohammed Alawad, M. Todd Young, John Gounley, Noah Schaefferkoetter, Hong Jun Yoon, Xiao-Cheng Wu, Eric B. Durbin, Jennifer Doherty, Antoinette Stroup, Linda Coyle, and Georgia Tourassi. 2021. Limitations of Transformers on Clinical Text Classification. *IEEE Journal of Biomedical and Health Informatics* 25, 9 (2021), 3596–3607. https://doi.org/10.1109/JBHI.2021.3062322

[9] Russell Gersten, Scott Baker, Timothy Shanahan, Sylvia Linan-Thompson, Penelope Collins, and Robin Scarcella. 2007. Effective Literacy and English Language Instruction for English Learners in the Elementary Grades. IES Practice Guide. NCEE 2007-4011. *What Works Clearinghouse* (01 2007).

[10] Edita Grolman, Andrey Finkelshtein, Rami Puzis, Asaf Shabtai, Gershon Celniker, Ziv Katzir, and Liron Rosenfeld. 2018. Transfer Learning for User Action Identication in Mobile Apps via Encrypted Trafc Analysis. *IEEE Intelligent Systems* 33, 2 (2018), 40–53. https://doi.org/10.1109/MIS.2018.111145120

[11] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A Survey on Contrastive Self-Supervised Learning. *Technologies* 9, 1 (2021). https://doi.org/10.3390/technologies9010002

[12] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[13] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

[14] Sinno Jialin Pan and Qiang Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (2010), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

[15] Makoto Sano, Doris Baker, Marlen Collazo, Nancy Le, and Akihito Kamata. 2020. Measuring the Expressive Language and Vocabulary of Latino English Learners Using Hand Transcribed Speech Data and Automated Scoring. *International Journal of Intelligent Technologies and Applied Statistics* 13, 3 (10 2020), 229-258 pages. https://doi.org/10.6148/IJITAS.202009_13(3).0003

[16] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605. http://jmlr.org/papers/v9/vandermaaten08a.html

[17] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. WELFake: Word Embedding Over Linguistic Features for Fake News Detection. *IEEE Transactions on Computational Social Systems* 8, 4 (2021), 881–893. https://doi.org/10.1109/TCSS.2021.3068519

[18] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3, 1 (2016), 1–40.