Working Paper

Bag-of-Words Algorithms Can Supplement Transformer Sequence Classification & Improve Model Interpretability

Christian Johnson and William Marcellino

RAND National Security Research Division

WR-A1719-1
January 2022

RAND
CORPORATION

For more information on this publication, visit www.rand.org/pubs/working_papers/WRA1719-1.html.

Published by the RAND Corporation, Santa Monica, Calif.

About This Working Paper

Classifying documents at scale using algorithms is an important sub-area within machine learning. Although current generation transformer models perform extremely well on many natural language tasks such as document classification, they struggle with computing and memory requirements on long sequences, and often require significant amounts of computing power to train. We describe a simple method of improving performance on the problem of classifying sequences of text by concatenating the hidden state of a BERT-based transformer model with a dictionary-based bag-of-words model. The hybrid models that result outperform the transformer models by varying margins, while adding trivial amounts of compute requirements and boosting model interpretability. Just as importantly, we show that this hybrid approach can improve interpretability of models.

Better performing and more interpretable text classification models are important across a range of applications but has particular significance for national security. Quickly and accurately detecting malign information campaigns, extremist recruitment content, or conspiracy theories circulated over social media serves national security interests. Additionally, understanding how these antisocial messages function can inform responses. While this paper is primarily written for a technical audience familiar with machine learning and natural language processing, it should be of interest within the operations in information environment (OIE) community.

National Security Research Division

This research was sponsored by the Office of the Secretary of Defense and conducted within the International Security Defense Policy (ISDP) Center of the RAND National Security Research Division (NSRD), which operates the RAND National Defense Research Institute (NDRI), a federally funded research and development center (FFRDC) sponsored by the Office of the Secretary of Defense, the Joint Staff, the Unified Combatant Commands, the Navy, the Marine Corps, the defense agencies, and the defense intelligence enterprise.

For more information on the RAND ISDP Center, see www.rand.org/nsrd/isdp or contact the director (contact information is provided on the webpage).

Acknowledgments

The authors would like to thank Pete Schirmer for his thoughtful review and helpful comments.

Abstract

Although transformer models perform extremely well on many natural language tasks, they may struggle with computing and memory requirements on long sequences, and often require significant amounts of computing power to train. Such models also lack interpretability. We describe a simple method of improving performance on the problem of classifying sequences of text by concatenating the hidden state of a BERT-based transformer model with a dictionary-based bag-of-words model. The hybrid models that result outperform the transformer models by varying margins, while adding trivial amounts of compute requirements and boosting model interpretability.

Summary

With the growth of the internet and a desire for data-driven solutions, policy analysis has increasingly relied on text mining to generate insights and recommendations. The RAND Corporation, for instance, has applied text analysis for identifying election interference on social media, detecting linguistic differences between conspiracy theories, and determining state-sponsored narratives around COVID-19, among other initiatives. Robust, interpretable models of language are critical to these analyses.

Language models based on the Transformer architecture have ushered in a new era in natural language processing (NLP) and have quickly become the *de facto* standard method for many general-purpose NLP tasks, from language translation to question-answering. However, Transformer models are notorious for their intensive computing requirements and sometimes difficult-to-interpret results, both of which originate from the Transformer's highly complex deep neural network architecture, which often utilizes millions or billions of parameters. Bag-of-words (BoW) models – models where the ordering of words is irrelevant, and all that matters is the set of words that are used and/or their occurrence frequency – are far simpler and computationally cheaper, albeit at the cost of substantially lower performance compared to more advanced Transformer models. Here we argue that simple BoW models, used in conjunction with deep Transformer models, can lead both to boosted performance and improved interpretability. This approach also is feasible in compute-constrained environments, e.g. without the use of GPU arrays.

We demonstrate this in two different ways: first, we describe an experiment where the addition of a BoW model dramatically improves sequence classification performance compared to a Transformer alone; and second, we show how the classification errors of a moderately inaccurate BoW model yields orthogonal insights to the classification successes of a pure Transformer model. Our results, while not groundbreaking, demonstrate the value of considering a suite of models when attempting to derive insights from data, instead of simply selecting the best-performing model, particularly in computing-constrained environments and where results must be translated into actionable policy.

Contents

Figures and Tables

Figures

Tables

1. Introduction

With the growth of the internet and a desire for data-driven solutions, policy analysis has increasingly relied on text mining to generate insights and recommendations. The RAND Corporation, for instance, has applied text analysis for identifying election interference on social media, detecting linguistic differences between conspiracy theories, and determining state-sponsored narratives around COVID-19, among other initiatives. Robust, interpretable models of language are critical to these analyses.

Language models based on the Transformer architecture (Vaswani et al. 2017) have ushered in a new era in natural language processing (NLP) and have quickly become the *de facto* standard method for many general-purpose NLP tasks, from language translation to question-answering. The attention mechanism that Transformer models employ allows them to deal with relatively long sequences of text in an efficient way, especially compared to older recurrent neural network architectures like seq2seq (Sutskever, Vinyals, and Le 2014). However, Transformer models are notorious for their intensive computing requirements and sometimes difficult-to-interpret results, both of which originate from the Transformer's highly complex deep neural network architecture, which often utilizes millions or billions of parameters.

Bag-of-words (BoW) models – models where the ordering of words is irrelevant, and all that matters is the set of words that are used and/or their occurrence frequency – are far simpler and computationally cheaper. Several BoW models that rely on pre-computed dictionaries for sentiment and sociocultural features of language have been developed for analysis, such as Linguistic Inquiry and Word Count (LIWC) (Pennebaker and Francis 1999), MoralStrength (Araque, Gatti, and Kalimeri 2020), or SEANCE (Crossley, Kyle, and McNamara 2017). The tradeoff of using this class of models, of course, is substantially lower performance compared to more advanced Transformer models. Biggiogera et al., for example, directly compared a model relying on LIWC to one built around Google's BERT[1] model (Devlin et al. 2019) for the task of predicting conflict in relationships, finding the BERT-based model to be superior (Biggiogera et al. 2021).

The goal of this paper is to argue that simple BoW models, used in conjunction with deep Transformer models, can lead both to boosted performance and improved interpretability. This approach also is feasible in compute-constrained environments, e.g. without the use of GPU arrays. In this paper we employ a dictionary-based model analogous to LIWC, which we refer to as *stance*, in our analysis.

By *stance* we mean the attitudinal dimension of language, for example affect, certainty, social relations--stance is part of the pragmatic function of language, the complement of

[1] BERT stands for Bidirectional Encoder Representations from Transformers

semantics (Kavanagh et al. 2019). Our model uses a taxonomy of 119 stance variables originally developed at Carnegie Mellon University to capture rhetorical and pragmatic effects in text (Ringler, Klebanov, and Kaufer 2018; Wetzel et al. 2021). This stance model is a general purpose one, well suited to a wide range of pre-planned texts, and general enough to have been ported over to Modern Standard Arabic successfully for clustering the Central Intelligence Agency's Bin-Laden archive (Bellasio et al. 2021). All documents in our datasets were transformed into vectors within that 119-dimension rhetorical space. These vectors tend to be relatively sparse, because most sentences do not contain any particular language stance, but become less sparse as the length of the document increases. For documents with about 20 words, about 10 stance values are typically nonzero, while documents with 100 words typically have about 25 nonzero stances.[2] This can be contrasted with the embeddings produced by a Transformer model, which are dense (nearly all values are nonzero). The stance model captures the pragmatic half of the hybrid model (a Transformer model captures the semantic half) and is human interpretable: features such as "fear," "uncertainty." or "spatial relations" make sense to humans.

Hybrid architectures combining machine learning and deep learning methods are not new. Because each has affordances and constraints, hybrid approaches can compensate for complementary weaknesses and improve model performance across a wide range of applications including sentiment classification (Salur and Aydin 2020), recommendation systems (Huang et al. 2019), and modeling the spread of infectious diseases (Chew et al. 2021). Where our hybrid approach is distinct is in leveraging a human theory of language features. This broad, general dictionary used not only improves classification performance in. many applications, but perhaps more importantly improves interpretability of models through a sparse taxonomy of language moves that are human readable.

As mentioned above, for the broad problem of analyzing text itself, rather than focusing solely on the best model for predictions, we believe both types of models can complement one another. We demonstrate this in two different ways: first, we describe an experiment where the addition of a BoW model dramatically improves sequence classification performance compared to a Transformer alone; and second, we show how the classification errors of a moderately inaccurate BoW model yields orthogonal insights to the classification successes of a pure Transformer model. Our results, while not groundbreaking, demonstrate the value of considering a suite of models when attempting to derive insights from data, instead of simply selecting the best-performing model, particularly in computing-constrained environments and where results must be translated into actionable policy.

[2] The fraction of nonzero stances tends to increase roughly with the square root of the document length.

2. Performance Improvements

Training a Transformer model to perform sequence classification can be done in one of two main ways: the model can be trained directly through backpropagation, or the sequence can be fed into pretrained Transformer whose hidden states are then used as training data for a simpler algorithm. The latter method is generally more appropriate for cases with limited amounts of training data or compute power, where fine-tuning the Transformer network is impractical, and is the method that we use here.

We observed a somewhat surprising result in (Marcellino et al. 2020) where a Transformer-based model was used to classify a dataset of known Russian trolls on Twitter. It was found that a direct concatenation of the stance vectors with the hidden state of a BERT-based model, fed through a logistic regression classifier, significantly boosted performance in some cases. Inspired by that result (which we reproduce below), we set out to investigate two other datasets to see if a similar boost in performance would occur. We refer to this kind of model that combines a BoW representation of text with a deep neural network embedding as a *hybrid* model below.

As a benchmark transformer model, we use the base DistilRoBERTa model from the `sentence-transformer` library (Reimers and Gurevych 2019). This model is a trained to generate a vector space of short text sequences, such that similar sentences are placed nearby one another in the space. Training is performed by digesting two separate sequences of text and training a 'Siamese network' on a label associated with the closeness of the two sequences. One half of the Siamese network is then removed, and an input sentence can be encoded by observing the final hidden state in the network. The result is a vector of length 768.

We build a simple classifier on the DistilRoBERTa and stance data by training a logistic regression model on top of either the DistilRoBERTa vector, stance vector, or the concatenation of the two. This straightforward implementation of a classifier is well-suited for scenarios with limited computing capabilities, as no fine-tuning or backpropagation is required for the DistilRoBERTa model and generating stance vectors scales linearly with the amount of text required. Such a model can be trained and applied even without the use of a GPU. The embeddings of the DistilRoBERTa model (which are essentially normally distributed around zero) and the stance vectors (which range from 0 to 1) are scaled differently, but a logistic regression model fits each parameter separately, so no rescaling was required before training. The logistic regression model we used in all cases applied a standard L2 regularization penalty to prevent overfitting with a value of 1, except for the Amazon dataset, as we describe below.

Table 1 shows the results of our 'out-of-the-box' implementations on three different text databases: Russian trolls on Twitter (Marcellino et al. 2020), ironic and non-ironic product reviews on Amazon (Filatova 2012), and the IMDB review dataset (Maas et al. 2011). The Russian troll dataset has 527 unique trolls, and 10,069 non-trolls, and is therefore highly imbalanced. The Amazon review dataset contains 437 ironic and 817 non-ironic reviews (1254 total samples), and the IMDB dataset contains 12,500 each of positive and negative reviews (for a total of 25,000). Note that while each dataset is nominally split into two distinct categories (troll/nontroll, ironic/nonironic, and positive/negative), they differ qualitatively. For example, negative movie reviews can range on a wide spectrum from mild to scathing, and irony can be expressed in multiple ways. On the other hand, because the Russian troll accounts were operated with a unified set of goals, we suspect that they may be more uniform in their language, which may explain the high performance we see below.

We split into training and testing sets with an 80/20 split ratio and train a logistic regression classifier on the associated vectors. Our classifier is built using `scikit-learn` (Pedregosa et al. 2011). All values shown are the Matthews correlation coefficient (Baldi et al. 2000) for the binary classification task, a performance metric that takes into account both true and false positives and negatives, and is well-suited to imbalanced datasets. Higher scores indicate better performance; we see that in all cases, the hybrid model performs the best (although gains in performance are varied). We interpret this variation in improvement as relating to the balance of pragmatic versus semantic content of the classified text. Russian trolls on the Internet seek to activate emotions and persuade, almost wholly pragmatic use of language, and thus a hybrid approach may add more value. On the other hand, Amazon product reviews, (especially non-ironic ones) are less rich in pragmatic content, and less likely to benefit from a hybrid approach.

We note that the Amazon dataset is relatively small (only 1207 total instances), which is probably why we see the strongest evidence of overfitting on it of our samples. In fact, the default regularization appeared to be insufficient on the Amazon dataset, so we used an L2 parameter of 0.07, which was the strongest regularization that still achieved reasonable performance. Still, even with this regularization, the hybrid model continues to perform best, suggesting that nontrivial information is contained in the two different datasets.

Also interesting to note is that the entries in the Russian Trolls data, which consists of concatenated tweets, are generally much longer (average of about 1000 words) than the IMDB (about 230) and Amazon (about 90) entries. Transformer models are notoriously difficult to train on longer pieces of text, while BoW models should perform best on long text. Indeed, we see the biggest gain in performance for the hybrid models on the Russian Trolls data, which suggests that long text sequences may be on of the best uses of hybrid modeling approaches.

Table 1. Classification Performance

Dataset	BoW Stance		DistilRoBERTa		Hybrid	
	Train	Test	Train	Test	Train	Test
Russian trolls	0.933	0.936	0.953	0.943	0.995	**0.992**
Amazon ironic reviews	0.545	0.443	0.492	0.493	0.659	**0.512**
IMDB	0.559	0.529	0.743	0.629	0.736	**0.659**

NOTE: Classification scores (Matthews correlation coefficient) for the three tasks with our three model paradigms (BoW, Transformer, and Hybrid). We find in all cases that the Hybrid model offers the best performance on the holdout (test) set.

3. Failure Modes Provide Insight

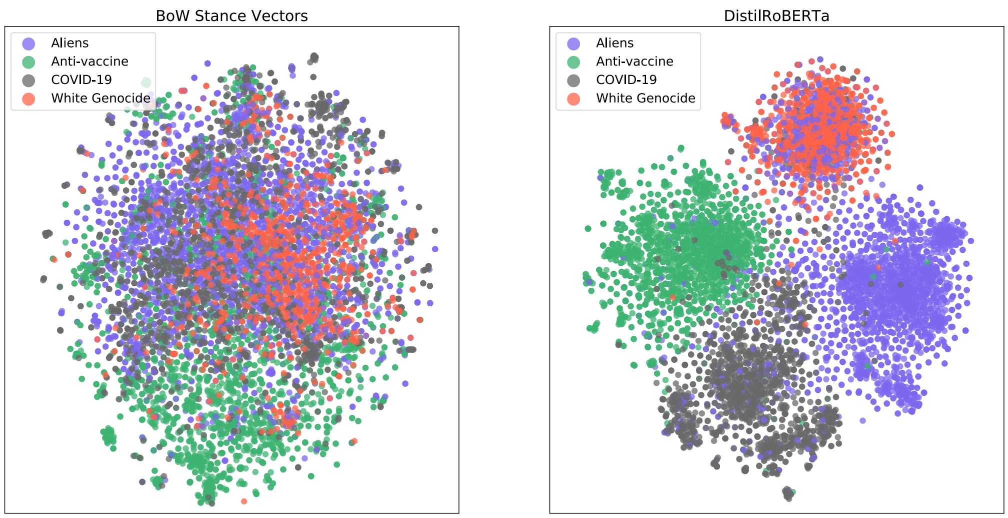
The primary goal when performing modeling is often to increase performance. A well-designed model, however, can often yield just as much insight (or even more) by observing its failures rather than its successes. We illustrate this principle by training and applying a DistilRoBERTa model and a BoW stance vector model to a dataset composed of comments associated with four different conspiracy theories: COVID-19, ‘white genocide', aliens, and anti-vaccination. More details about the dataset and the conspiracy theories themselves can be found in (Marcellino et al. 2021). The stance values are derived as before. For the DistilRoBERTa vectors, we use the embeddings produced by the Base v1 version of `sentence-transformers`. We then apply T-SNE dimensionality reduction (van der Maaten and Hinton 2008) to the resulting vectors and display the 2-dimensional projections in Figure 1.

Figure 1 shows that the DistilRoBERTa model is clearly superior to raw stance vectors when distinguishing between conspiracy theory language; not only are the clusters of comments clearly separable, we can even identify subcommunities within each conspiracy theory that are closely clustered together. This hierarchical mapping of conspiracy theories could be quite useful from a policymaker perspective, who may be interested in understanding how these communities change over time and interact with one another. Meanwhile, the clusters determined by the BoW stance data alone are rather indistinct, indicating that there is significant overlap in the type of language used by conspiracy theorists. Nevertheless, some conclusions are identifiable from the BoW stance model in Figure 1, namely, that anti-vaccine language is relatively distinct from other conspiracy theory language.

As with the Russian troll data, we next trained a logistic regression classifier (using one-versus-rest policy) on the BoW and DistilRoBERTa vectors, in order to determine how powerful each algorithm was for classifying sequences. We split the comments randomly into train (75%) and test (25%) sets, The results were unsuprising: the DistilRoBERTa model was highly accurate (>95%) on both training and test sets at identifying the conspiracy theory, while the BoW stance model was somewhat less accurate (87%). When trained on the combined DistilRoBERTa and BoW stance vectors, performance was barely changed from the DistilRoBERTa-alone model. The overall results are shown in Table 2.

At first glance, the results appear to indicate that the Transformer model is simply superior to the BoW stance model. But a closer look at the confusion matrices for our models (shown in Figure 2) demonstrates that the inaccuracies of the BoW stance model themselves provide insight: the model appears to confuse the Aliens and COVID-19 conspiracy theories significantly more than other theories. This indicates that these two conspiracy theories share several features of language outside of the semantic content. From Figure 1, one might assume that the Aliens conspiracy theory is relatively isolated; the fact that the two theories share many stance features might indicate that the two theories are potentially persuasive to each other's adherents.

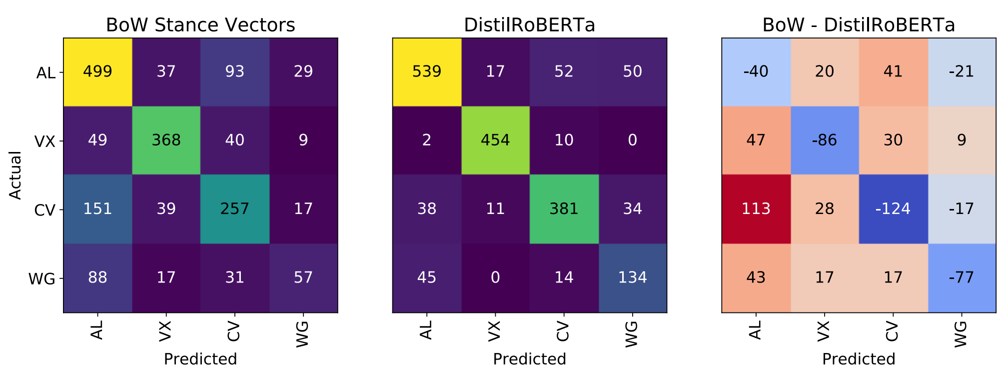
Figure 1. Conspiracy Theory Comments



SOURCE: RAND Analysis

NOTE: Comments associated with four conspiracy theories, mapped into a 2-dimensional space by the T-SNE algorithm. On the left, T-SNE is applied to the BoW stance vectors associated with each comment, while on the right, itis applied to the neural network hidden state in a DistilRoBERTa model. The Transformer is clearly more powerful at distinguishing between conspiracy theory language, even identifying sub-communities. The BoW stance model does not look at semantic content, which results in poorer distinguishing ability; however, some patters (anti-vaccine language is relatively distinct from other conspiracy theories) are clear.

Figure 2. Conspiracy Theory Confusion Matrices



SOURCE: RAND Analysis

NOTE: Confusion matrices (test sets only) for Logistic Regression classification algorithm trained on the BoW stance vectors and the DistilRoBERTa hidden states. In the right panel is the difference between the two confusion matrices, for clarity. As before, the DistilRoBERTa model is more accurate than the BoW stance model, but the inaccuracy in distinguishing Aliens from COVID-19 comments is telling nonetheless - an insight that would be missed by a Transformer-only model.

Discussion

Understanding text corpora is important for a variety of problems across public policy domains, such as developing taxonomies of conspiracy theories or identifying foreign propaganda. The results that we present here suggest that the use of shallow, interpretable BoW models can improve understanding of text corpora when used in combination with deep Transformer language models. We note that not only do these models scale well as text length increases (unlike most Transformer models), they improve performance of Transformer classification algorithms, in some cases significantly. This is interesting, because the two models digest the same data, meaning they must be extracting different data from the same sequences of text. This potentially points towards a promising direction for future Transformer architectures; we leave this question for future work. As we also showed, improved models are not necessarily always ideal -- models should be expected to fail when presented with certain data. Model failure, therefore, can be a clue that identifies interesting patterns in data that would not be found otherwise. These methods could be useful to supplement future text analyses that require both power and interpretability, such as those done in a policymaking context.

Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
BoW	Bag of Words
LIWC	Linguistic Inquiry and Word Count
NLP	Natural Language Processing

References

Araque, Oscar, Lorenzo Gatti, and Kyriaki Kalimeri. 2020. "MoralStrength: Exploiting a Moral Lexicon and Embedding Similarity for Moral Foundations Prediction." *Knowledge-Based Systems* 191 (March): 105184. https://doi.org/10.1016/j.knosys.2019.105184.

Baldi, Pierre, Søren Brunak, Yves Chauvin, Claus A. F. Andersen, and Henrik Nielsen. 2000. "Assessing the Accuracy of Prediction Algorithms for Classification: An Overview." *Bioinformatics* 16 (5): 412–24. https://doi.org/10.1093/bioinformatics/16.5.412.

Bellasio, Jacopo, Sarah Grand-Clement, Shazan Iqbal, William Marcellino, Alice Lynch, Yousuf Abdelfatah, Tor Richardson-Golinski, Kate Cox, and Giacomo Persi Paoli. 2021. "Insights from the Bin Laden Archive: Inventory of Research and Knowledge and Initial Assessment and Characterisation of the Bin Laden Archive," May. https://www.rand.org/pubs/research_reports/RRA109-1.html.

Biggiogera, Jacopo, George Boateng, Peter Hilpert, Matthew Vowels, Guy Bodenmann, Mona Neysari, Fridtjof Nussbeck, and Tobias Kowatsch. 2021. "BERT Meets LIWC: Exploring State-of-the-Art Language Models for Predicting Communication Behavior in Couples' Conflict Interactions." *ArXiv:2106.01536 [Cs]*, June. http://arxiv.org/abs/2106.01536.

Chew, Alvin Wei Ze, Yue Pan, Ying Wang, and Limao Zhang. 2021. "Hybrid Deep Learning of Social Media Big Data for Predicting the Evolution of COVID-19 Transmission." *Knowledge-Based Systems* 233 (December): 107417. https://doi.org/10.1016/j.knosys.2021.107417.

Crossley, Scott A., Kristopher Kyle, and Danielle S. McNamara. 2017. "Sentiment Analysis and Social Cognition Engine (SEANCE): An Automatic Tool for Sentiment, Social Cognition, and Social-Order Analysis." *Behavior Research Methods* 49 (3): 803–21. https://doi.org/10.3758/s13428-016-0743-z.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *ArXiv:1810.04805 [Cs]*, May. http://arxiv.org/abs/1810.04805.

Filatova, Elena. 2012. "Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 392–98. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/661_Paper.pdf.

Huang, Zhenhua, Chang Yu, Juan Ni, Hai Liu, Chun Zeng, and Yong Tang. 2019. "An Efficient Hybrid Recommendation Model With Deep Neural Networks." *IEEE Access* 7: 137900–912. https://doi.org/10.1109/ACCESS.2019.2929789.

Kavanagh, Jennifer, William Marcellino, Jonathan S. Blake, Shawn Smith, Steven Davenport, and Mahlet Gizaw. 2019. "News in a Digital Age: Comparing the Presentation of News Information over Time and Across Media Platforms," May. https://www.rand.org/pubs/research_reports/RR2960.html.

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. "Learning Word Vectors for Sentiment Analysis." In *Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies, 142–50. Portland, Oregon, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/P11-1015.
Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research* 9 (11): 2579–2605.
Marcellino, William, Todd C. Helmus, Joshua Kerrigan, Hillary Reininger, Rouslan I. Karimov, and Rebecca Ann Lawrence. 2021. "Detecting Conspiracy Theories on Social Media: Improving Machine Learning to Detect and Understand Online Conspiracy Theories." RAND Corporation. https://doi.org/10.7249/RR-A676-1.
Marcellino, William, Christian Johnson, Marek N. Posard, and Todd C. Helmus. 2020. "Foreign Interference in the 2020 Election: Tools for Detecting Online Election Interference," October. https://www.rand.org/pubs/research_reports/RRA704-2.html.
Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.
Pennebaker, James W., and Martha E. Francis. 1999. *Linguistic Inquiry and Word Count*. Lawrence Erlbaum Associates, Incorporated.
Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084.
Ringler, Hannah, Beata Beigman Klebanov, and David Kaufer. 2018. "Placing Writing Tasks in Local and Global Contexts: The Case of Argumentative Writing." *Journal of Writing Analytics* 2: 34–77. https://doi.org/10.37514/JWA-J.2018.2.1.03.
Salur, Mehmet Umut, and Ilhan Aydin. 2020. "A Novel Hybrid Deep Learning Model for Sentiment Classification." *IEEE Access* 8: 58080–93. https://doi.org/10.1109/ACCESS.2020.2982538.
Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. "Sequence to Sequence Learning with Neural Networks." In *Advances in Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger. Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." *ArXiv:1706.03762 [Cs]*, December. http://arxiv.org/abs/1706.03762.
Wetzel, Danielle, David Brown, Necia Werner, Suguru Ishizaki, and David Kaufer. 2021. "Computer-Assisted Rhetorical Analysis: Instructional Design and Formative Assessment Using DocuScope." *The Journal of Writing Analytics* 5. https://doi.org/10.37514/JWA-J.2021.5.1.09.