Classification performance thresholds for BERT-based models on COVID-19 Twitter misinformation

Johnattan Ontiveros, Robyn Correll Carlyle, Anika Puri, Sagar Kumar, Alexander Tregub, Caroline Nitirahardjo, Evelynne Morgan, Brendan Lawler, Eliza Aimone, Dr. Helen Piontkivska, Dr. Maimuna Majumder

Corresponding author: Dr. Maimuna Majumder, maimuna.majumder@childrens.harvard.edu, 617-355-6000

Affiliations:

JO: Computational Health Informatics Program (CHIP), Boston Children's Hospital, Boston, MA, USA
CompEpi Dispersed Volunteer Research Network (DVRN), Boston, MA, USA
Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA, USA, johnonti@mit.edu
RCC: Vaccinate Your Family, Washington, DC, USA, robyn@vaccinateyourfamily.org
AP: Computational Health Informatics Program (CHIP), Boston Children's Hospital, Boston, MA, USA
CompEpi Dispersed Volunteer Research Network (DVRN), Boston, MA, USA, anpuri@mit.edu
SK: Network Science Institute at Northeastern University, Boston, MA, USA, kumar.sag@northeastern.edu
AT: Department of Mathematical Sciences, Kent State University, Kent, OH, USA, atregub@kent.edu
CN: Department of Biological Sciences, Kent State University, Kent, OH, USA, cnitirah@kent.edu
EM: Department of Biological Sciences, Kent State University, Kent, OH, USA, emorga26@kent.edu
BL: Computational Health Informatics Program (CHIP), Boston Children's Hospital, Boston, MA
Department of Pediatrics, Harvard Medical School, Boston, MA, USA
CompEpi Dispersed Volunteer Research Network (DVRN), Boston, MA, USA, brendanlawler2020@gmail.com
EA: Department of Biological Sciences, Kent State University, Kent, OH, USA, sdohmlo@kent.edu
HP: Department of Biological Sciences, Kent State University, Kent, OH, USA
CompEpi Dispersed Volunteer Research Network (DVRN), Boston, MA, USA, opiontki@kent.edu
MM: Computational Health Informatics Program (CHIP), Boston Children's Hospital, Boston, MA, USA
Department of Pediatrics, Harvard Medical School, Boston, MA, USA
CompEpi Dispersed Volunteer Research Network (DVRN), Boston, MA, USA
maimuna.majumder@childrens.harvard.edu

Funding information:

JO, AP, CN, EM, EA, and HP were supported in part by grant SES2230083 from the National Science Foundation. SK was supported in part by grant SES2200228 from the National Science Foundation. The CompEpi Dispersed Volunteer Research Network is sponsored in part by grant R35GM146974 from the National Institute of General Medical Sciences, National Institutes of Health. The funding sources had no involvement in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; or in the decision to submit the paper for publication.

Competing interest declaration:

We know of no conflicts of interest associated with this publication, and there has been no financial support associated with this research that could have affected its outcome.

Ethical approval statement or statement of informed consent for case studies:

Not applicable

Trial registration details:

Not applicable

Main text:

Applications of Artificial intelligence (AI) skyrocketed during the COVID-19 era. Among the AI subfields gaining momentum is natural language processing (NLP), a methodology that can analyze large amounts of text, making it useful for evaluating public attitudes, detecting outbreaks, and identifying misinformation — particularly on social media platforms like Twitter.[1–3]

A branch of NLP known as sentiment classification aims to identify authors' views or opinions from a text-based corpus.[4] One of the best-performing models for analyzing the Twitter corpus is Bi-directional Encoder Representations from Transformers (BERT).[4,5] Pre-trained on a large corpus of general text, BERT is able to infer "context" from text, enabling it to outperform simpler models that leverage the use of keywords to infer the meaning of text.[5]

When classifying sentiment on niche topics, BERT can be fine-tuned with hand-labeled, subject-specific text — a step that is generally accepted as necessary for improving accuracy.[6] However, the process of producing hand-labeled data can have unforeseen downstream effects, and it is unclear how the amount of data used for fine-tuning can impact model performance.

Recently, researchers developed a binary-class BERT model by hand-labeling the sentiment of 15000 tweets containing COVID-19 vaccine-related keywords (Supplement) and fine-tuning BERT to identify "Anti-Vax" misinformation.[7] Their model achieved a high test-set accuracy of 98% using a 9547-tweet corpus of training data, making their approach an ideal experimental case for gauging how the quantity and selection of training data might impact BERT's performance on a specific topic.

To explore these effects, full text of the sentiment-labeled tweet-ids from were collected using Twitter Hydrator.[7,8] Then, the distilbert-base-uncased model[9] — a type of BERT model (Supplement) — was trained upon progressively increasing increments of the total 15k-tweet, hand-labeled Anti-Vax dataset (10% [N=1150], 20% [N=2300], 30% [N=3450], etc.) and evaluated on a withheld test set of 10%. For each increment, the data was randomly sampled three times and trained on the data for 20 epochs (i.e., full passes through the data). The cross-sample mean of the test set's best F1-scores was then computed for any given increment. The intention of this design was to saturate the models with the training data, and then record the highest achieved test scores at any given training size. As the training *N* increased, we expected test scores to improve and eventually flatten as increased size yielded smaller returns in accuracy.

With the distilbert-base-uncased model, the curve in Figure 1a flattened rapidly even when small fractions of the training data were used. This may be attributable to two things: [1] the wide adaptability of large language models (LLM) like BERT or [2] the easy separability of classes from the selected keywords.

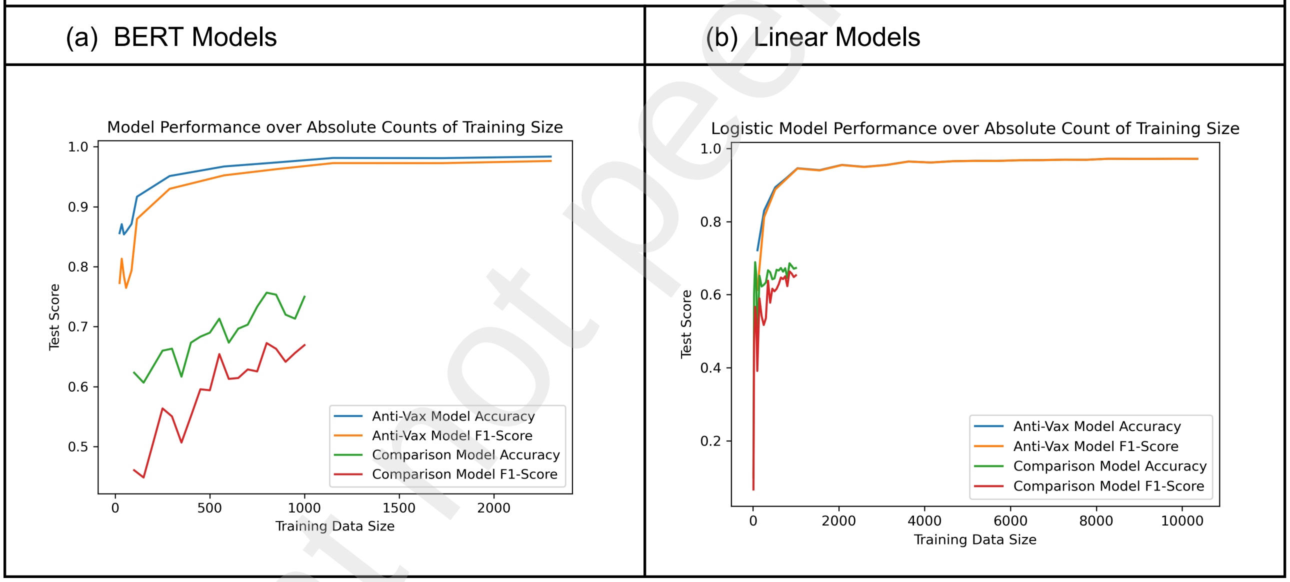
To explore the first possibility, we attempted the above experimental procedure with a simple logistic linear classifier, using the Word2Vec algorithm to convert the tweets into numerical form. The results from Figure 1b show similarly rapid conversion results as in Figure 1a without the advantages of an LLM.

To explore the second possibility, we trained the distilbert-base-uncased model using a secondary comparison dataset with topically-similar, but more generalized, keywords (Supplement). We did this by collecting 1000 tweets from February 2022 using the Twitter Application Programming Interface, and hand-labeling them as having anti-vaccine, neutral, or pro-vaccine sentiment. To mirror the binarized structure of the Anti-Vax dataset, labels for pro-vaccine and neutral were grouped into a "not anti-vaccine" class. Following the aforementioned experimental procedure, we likewise trained and tested the comparison model using incremental sets of hand-labeled data (10% [N=90], 20% [N=180], 30% [N=270], etc.). The results were then compared to the Anti-Vax dataset (Figure 1).

The model using the comparison dataset yielded substantially lower F1 and accuracy scores than the models trained on the Anti-Vax data with similarly sized training sets of <1000 tweets. For *N* <1000, the highest scores achieved in the Anti-Vax data were 98% for accuracy and 97% for F1, while scores for the comparison dataset did not surpass 76% for accuracy and 67% for F1. This suggests that a higher accuracy can be achieved with smaller hand-labeled Twitter datasets by using a narrower set of keywords. However, LLMs like BERT are prone to achieving high scores by taking syntactic short-cuts that break down when tasked with more complex classification challenges such as an evolving Twitter lexicon.[10]

Despite this important finding, our analysis had several limitations. The datasets considered were specifically related to COVID-19 vaccine misinformation and anti-vaccine sentiment; thus, our findings may not be transferable to other niche classification analyses. Likewise, BERT models in this analysis were trained on binary classification labels; performance of BERT models trained using datasets with more than two labels may require different amounts of data to achieve similar results. Further, while we focus on Twitter data in this work, we expect that our findings also apply to other platforms such as Reddit, where future work can expand in response to ongoing shifts across the social media landscape.

Figure 1: Model performance over absolute training size



(a) BERT Models

(b) Linear Models

References:

1. Liu S, Liu J. Public attitudes toward COVID-19 vaccines on English-language Twitter: A sentiment analysis. *Vaccine* [Internet]. 2021 Sep 15 [cited 2023 Apr 20];**39**(39):5499–505. Available from: https://www.sciencedirect.com/science/article/pii/S0264410X21011063
2. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *J Am Med Inform Assoc* [Internet]. 2008 Mar 1 [cited 2023 Apr 20];**15**(2):150–7. Available from: https://doi.org/10.1197/jamia.M2544
3. Nistor A, Zadobrischi E. The Influence of Fake News on Social Media: Analysis and Verification of Web Content during the COVID-19 Pandemic by Advanced Machine Learning Methods and Natural Language Processing. *Sustainability* [Internet]. 2022 Jan [cited 2023 Apr 20];**14**(17):10466. Available from: https://www.mdpi.com/2071-1050/14/17/10466
4. Zunic A, Corcoran P, Spasic I. Sentiment Analysis in Health and Well-Being: Systematic Review. *JMIR Med Inform*. 2020 Jan 28;**8**(1):e16023.

5. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv; 2019 [cited 2023 Apr 20]. Available from: http://arxiv.org/abs/1810.04805
6. Rogers A, Kovaleva O, Rumshisky A. A Primer in BERTology: What We Know About How BERT Works. *Trans Assoc Comput Linguist* [Internet]. 2021 Jan 1 [cited 2023 Jun 5];**8**:842–66. Available from: https://doi.org/10.1162/tacl_a_00349
7. Hayawi K, Shahriar S, Serhani M, Taleb I, Mathew S. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health* [Internet]. 2021 Dec 7 [cited 2023 Jan 20];**203**:23–30. Available from: https://doi.org/10.1016/j.puhe.2021.11.022.
8. Hydrator [Internet]. Documenting the Now; 2023 [cited 2023 Jun 23]. Available from: https://github.com/DocNow/hydrator
9. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [Internet]. arXiv; 2020 [cited 2023 Jun 19]. Available from: http://arxiv.org/abs/1910.01108
10. McCoy RT, Pavlick E, Linzen T. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference [Internet]. arXiv; 2019 [cited 2023 Jun 14]. Available from: http://arxiv.org/abs/1902.01007

SUPPLEMENTARY MATERIALS

Code Repository: https://github.com/JohnOnt/Classification-Thresholds-BERT-Misinfo

1.) Keywords Used to Create Data Sets and their Occurrences

The table below specifies the keywords used by the reference study ("Anti-Vax") to select tweets for hand-labeling, and the keywords used to formulate the comparison data set. Notably, the keywords used for the collection of the comparison data span a wider set of possible matches by accepting tweets that contain any word or phrase that begins with "covid", "coronav", etc. Because of these wider sets of possible matches that refer to Covid-19 and vaccines as a whole, this comparison data set is representative of a more generalized set of keywords on the topic.

Table 1: Keywords Used to Create Dataset and their Occurrences

	Keyword (uncased)	**Count (Not "Anti-Vaccine")**	**Count ("Anti-Vaccine")**
15k Anti-Vax Data	All Tweets	7628	3875
	Vaccine	6765	3515
	Pfizer	884	462
	Moderna	630	176
	Astrazeneca	180	105
	Sputnik	11	5
	Sinopharm	25	3
1k Comparison Data	All Tweets	595	404
	Covid*	437	256
	Coronav*	7	2
	Vaccin*	347	168
	Vax*	524	344

** Indicates that words that extend the substring are matched; for example, Covid and Covid-19 are matched, but Coron is not.*

2.) Selection and Training of BERT Model

The original introduction of the BERT architecture highlighted two commonly used versions of the architecture, BERT-base and BERT-large, differentiated by their size (number layers and self-attention heads).[1] An increased number of parameters within a given model architecture generally increases the amount of "compute," or resources, required to pre-train and fine-tune the model.[1,2] For the purposes of this paper a large number of iterations of fine-tuning were required with relatively little compute resources available, leading us to choosing the "BERT-base-uncased" model (110M parameters) as opposed to "BERT-large-uncased" (340M parameters).

In the referenced “Anti-Vax” misinformation detection paper, the “BERT-large-uncased” model is used,[3] but our replications with the “BERT-base-uncased” model successfully reproduce the reported scores in the paper. All reported BERT-related scores in the main text are therefore produced using “BERT-base.”

3.) Fine-Tuning Approach

The fine tuning process runs the BERT model through 20 epochs of the training data three times, and takes the cross-average of its best validation scores achieved across those 20 epochs. The selection of 20 epochs was set as an upper bound estimate for when the model would have been sufficiently saturated by the training data, allowing us to choose the model that best performed on the validation set. The upper bound of 20 was determined through individual testing for both the “Anti-Vax” data and comparison data set.

References

1. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Internet]. arXiv; 2019 [cited 2023 May 29]. Available from: http://arxiv.org/abs/1810.04805
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language Models are Few-Shot Learners [Internet]. arXiv; 2020 [cited 2023 Jun 1]. Available from: http://arxiv.org/abs/2005.14165
3. Hayawi K, Shahriar S, Serhani MA, Taleb I, Mathew SS. ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. Public Health. 2022 Feb;**203**:23–30.