On the Gap between Adoption and Understanding in NLP

Federico Bianchi
Bocconi University
Milano, Italy
f.bianchi@unibocconi.it

Dirk Hovy
Bocconi University
Milano, Italy
dirk.hovy@unibocconi.it

Abstract

There are some issues with current research trends in NLP that can hamper the free development of scientific research. We identify five of particular concern: 1) the early adoption of methods without sufficient understanding or analysis; 2) the preference for computational methods regardless of risks associated with their limitations; 3) the resulting bias in the papers we publish; 4) the impossibility of re-running some experiments due to their cost; 5) the dangers of unexplainable methods. If these issues are not addressed, we risk a loss of reproducibility, reputability, and subsequently public trust in our field. In this position paper, we outline each of these points and suggest ways forward.

1 Early Adoption

When BERT (Devlin et al., 2019) was introduced in 2019, it revolutionized Natural Language Processing (NLP), showing impressive capabilities on many tasks in various languages (Nozza et al., 2020). However, papers soon highlighted its limits (Ettinger, 2020; Bender and Koller, 2020) and identified issues with bias (Nozza et al., 2021), i.e., that contextual models show different performances for different genders and ethnic groups (Zhang et al., 2020). While Rogers et al. (2020a)'s "BERTology" paper outlines what we know about BERT, much remains unexplored. We do not really know what BERT "understands", but, as Bender and Koller (2020) pointed out, we frequently overestimate its understanding capabilities. E.g., Ettinger (2020) shows that BERT has very low sensitivity to the concept of negation shows. This is not to say that we should stop using BERT; indeed, it is hard to imagine the field today without it. But it does illustrate the *gap between adoption and understanding* (GAU) of a new technology. As the field grows, this situation is likely to play out more frequently, as it is aided by various circumstances.

Early adoption of novel methods and their rigorous testing by the field are a strength of NLP. This approach has propelled insights through cascading waves of novel methods. However, the adoption of new technologies without full awareness of their potential and side effects is a risky proposition. In the 1950s and 1960s, a German pharmaceuticals company aggressively marketed a new drug called "Contergan" to treat sleeping problems as well as morning sickness during pregnancy. The drug had been classified as safe after extensive trials, and was widely prescribed. However, the trials had actually excluded pregnant women, so it was only after its approval that the effects of the drug's main component, thalidomide, became clear: thousands suffered severe birth defects and miscarriages (Criado Perez, 2019).

NLP is not a chemical product, and its effects are not as physically harmful. Using a model that later turns out to be overfitting is not in the same category as failing to protect people from bodily harm. However, it has other consequences. Say researcher A publishes a new result with method X, which becomes the number to beat. Many try and fail, not publishing their results or being ignored. Later, method X is found to be wrong. But the insights that were shelved or ignored in between are lost, unable to unseat method X's false supremacy. Maybe A retracts their initial result, but the damage is done. If others have built on X in the meantime, they will also be affected by its collapse. Imagine for a moment the ripple effects if a central paper like BERT turned out to be wrong. More likely, though, is that A, satisfied with the results of X, simply moves on. The barrier to new, better work remains in place, the faulty method is not identified as such, and instead a method heralded as revolutionary causes stagnation in the field.

Social psychology has struggled with this "win-

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3895–3901

ner's curse" (Ioannidis, 2008) of unreplicable results. The cause there was the prevalence of false positives that appeared significant, but were just lucky flukes (Ioannidis, 2005). Simmons et al. (2011) found that in many cases, a researcher's decision (conscious or subconscious) had influenced the results. While the causes are different, the resulting issue is the same in NLP.

We will see in the subsequent sections how, counter-intuitively, the GAU creates incentives to publish faster, focus on methods, and drop experiments that do not lead to state-of-the-art results.

A way forward We need to create an environment that makes negative findings and exploration of shortcomings possible, and preferably not just as afterthoughts. Workshops on negative results (Rogers et al., 2020b) and the stress-testing idea of build-it, break-it (Ettinger et al., 2017) are steps in the right direction.

2 Computational Papers

The method-driven nature of NLP makes it necessary to constantly explore new techniques, and the upsides are readily apparent. However, it has also tipped the balance of papers away from introspection and linguistically motivated studies. This development seems rooted in the statistical revolution that began in the 1990s, when method-driven papers outperformed theory-motivated ones. By now, it seems this attitude is very much ingrained in our community, where new models are appreciated more than purely linguistic results. To be fair, purely methodological papers are easier to evaluate objectively. This situation invites two questions, though: (i) Are modeling results more important than linguistics insights? And (ii) should computational papers be evaluated differently?

Norvig (2017) has remarked on the tension between rationalism, which wants models that can be understood, and empiricism, which wants models that work.[1] There is likely always going to be a pendulum swing between these extremes. But to make true progress, we do need both approaches.

While to date no survey has quantified the most popular methods in NLP research, it seems anecdotally that many of the accepted papers at top conferences in the field introduce novel models. Assuming that modeling results are important in our field, we need to understand if we are evaluating these papers in the correct way. This brings up the question of replicability: there has recently been a push to find ways to require authors to share their code and parameters, but this is not enough. Papers often fail to include the complete setup on which they base their models. Even the slightest difference in the setting can bring huge differences in the results: changing between CUDA versions can affect the results on GPU. Moreover, papers' repositories are often incomplete, failing to include the dependencies that would allow others to easy replicate experiments or, worse, containing only poorly documented Jupyter notebooks. Code is a fundamental component of science and should be regarded as such. Code should not just enable researchers to re-run a given experiment, but to apply the same method to other data sets. Bad code is not useful, and the cost of re-implementation, combined with the risk of not being able to get the original results, is often too high to justify the investment of researchers' time. Writing bad code should be akin to writing a bad paper; it does not necessarily make the research wrong, but it makes it less reproducible.

This situation brings up another question: Should authors be responsible for actively maintaining code? Once a new paper is published and the code released, reviewers might ask for comparisons. However, it is common to see repositories on GitHub that have many unanswered questions.

Methodological errors can slow research; as shown by Musgrave et al. (2020), due to methodological flaws in the experiments, the increase of performance in metric learning was wrongly reported in several papers. Moreover, a few hyperparameters can make a huge difference in the results of a given experiment (Ruffinelli et al., 2020). Well-documented methodological design and code make it easier to find bugs and experimental problems. Often systematic evaluations are run to compare different state-of-the-art methods and show that the results are sometimes on par with baselines (Dacrema et al., 2019; Errica et al., 2020).

A way forward Following Bender and Friedman (2018) on providing a data statement, we believe that a code statement similar to that offered by Mitchell et al. (2019) should be provided along with a paper. Code is part of the scientific contribution similar to the experiments we run to prove a hypothesis. While there can be different opinions

[1] The empiricist preference can be summed up with statistician George Box's aphorism, "All models are wrong, but some are useful."

on how to write good or bad code, writing documented code that is at least easy to use (with the use of high-level interfaces and convenient wrappers) can be required practice. Moreover, this can be of help in the context of systematic evaluation initiatives like code base that integrate multiple algorithms (Terragni et al., 2021): this process can help in reducing methodological errors in the evaluation. We can cite HuggingFace[2] as a notable example of a repository with good systematicity that has allowed many researchers to replicate most of the recent and popular research in the NLP field. Similarly, Sentence BERT (Reimers and Gurevych, 2019) - a method for sentence embeddings based on BERT - was released with a well organized and easy-to-use source code[3] that made it possible to use the models as a baseline or as the foundation of many other research works (Li et al., 2020; Ke et al., 2020; Zemlyanskiy et al., 2021; Bianchi et al., 2021c,b, inter alia), suggesting that this kind of release can be of great benefit to the community.

3 Publication Bias

Most researchers want to publish in A+ venues and Q1 journals as academic positions use publication output as a criterion for promotions. While quality *should* play into this assessment, it would be wrong to assume university committees are familiar enough with all venues and subfields in a discipline to make an accurate assessment of the relative value of each contribution. So the number of papers and citation count often trump other considerations, especially when publications give universities PR.

Proponents of "slow science" have argued for a shift away from this emphasis on quantity. But while laudable in theory, committing to slowness in practice does not align with the needs of especially junior researchers. They might well prefer less publication pressure, but they can likely not afford the trade-off between a theoretically desirable publication model and losing their job. Given all that, it would be unfair to put the responsibility for addressing the publication bias feeding the GAU on junior scholars. Furthermore, it is not clear what makes a paper worthy of an A+ venue. For NLP conferences, that decision is left to three reviewers and one area chair. Faced with an increasing number of papers to review, reviewers have found themselves with less and less time to make a well-rounded decision on each one. As a consequence, many good papers do not get published.

To keep up with publication demands, authors can 1) make their papers easier to judge based on a quick read or 2) find alternative venues. The former means focusing on a single, easily recognizable contribution—much easier for method papers (see also the previous section). The latter means many authors are now deciding to build the foundation of their publication record by publishing on ArXiv. There is nothing wrong with ArXiv itself. Thanks to it, researchers can make and share valuable contributions with other researchers online. However, ArXiv is a new, and changeable, venue. Keeping the current pace of NLP research, it is bound to also publish models that are biased. So publication records built on ArXiv are set on quicksand and might fall over: the GAU we described previously. While it is encouraging to see that we as a field subsequently work on reducing the bias of those models, we are still allowing anyone to use and deploy those biased models in the meanwhile, contributing to the GAU.

A way forward Short of changing the incentive structure, we can do more to strengthen the review process. Tutorials on reviewing (Cohen et al., 2020) and the implementation of their recommendations would go a long way toward ensuring that we maintain a high standard at a high volume.

4 Computationally Unobtainable

The previous sections have argued that the GAU is a consequence of a strong preference within the field for computational methods, amplified by existing publication bias. But even if all of the solutions we suggested were adopted, there would still be issues that affect reproducibility.

In a panel discussion at EurNLP 2019, Phil Blunsom correctly remarked that "[t]he future of NLP is not about bigger models, it's about bigger ideas."[4] Indeed, there are many arguments against simply making models bigger. However, there is a difference between having the possibility of running a bigger model and not needing it, and needing a bigger model but not having it.

Popular methods like BERT or GPT-3 are now impossible to develop without huge amounts of funding. Developing a new algorithm means run-

[2] huggingface.co

[3] https://github.com/UKPLab/sentence-transformers

[4] https://twitter.com/glorisonne/status/1182693114672271360

ning multiple experiments, multiple times, adjusting parameters and configurations. So there is not just the already prohibitive cost of pre-training one BERT model, but the cost of pre-training BERT dozens of times.

Asking authors to provide results by re-training models is too high a cost for many academic researchers. True, other fields have to invest much more money to run experiments. E.g., neurosciences require universities to buy ECG or MEG devices that can cost up to 2 million dollars. On the other hand, those devices are much more consistently reused than a single pre-trained model, so costs are much more distributed.

It is also unclear what constitutes a "bigger" model. Parameter sizes have grown exponentially for the last few years (Bender et al., 2021), and what was considered preposterously large five years ago is now pretty standard. But even old standards are becoming hard to match for universities. This situation creates an unbridgeable gap between industry and academia. Even within those two groups, there are rapidly emerging differences between the players. Only rich universities can afford to re-pre-train models from scratch. Demanding that all actors have access to the same resources would be unreasonable (however desirable). Industry players need to maintain a competitive edge, and we can hardly hope to address the inequality between national academic systems and their funding. But this reality creates a situation where reproducibility becomes impossible. If team A presents results from a model that only team A can train, then the rest of us need to take those results at face value. Whether we believe them or not becomes irrelevant. In addition to this problem, consider the environmental concerns generated by the training of large models (Strubell et al., 2019), and bigger does not necessarily equate better for reproducibility.

Lack of reproducibility, though, is a danger to scientific reliability. The fallout from the reproducibility crisis in social psychology (see Section 4) has tainted the reputation of the entire field. But it has also led to remarkable innovations. Studies now need to be pre-registered to prevent "fishing expeditions" to find the most "interesting" result. International research teams have organized trials to replicate famous experiments—often disproving canonical theories, but also re-opening avenues of research that had been wrongly foreclosed (Collaboration, 2015). And while NLP researchers generally favor reproducibility (Mieskes et al., 2019), we are not yet doing it. Fokkens et al. (2013) already identified five parameters for better reproducibility. A study by Belz et al. (2021), though, found that only 15.61% of 506 papers they checked were independently reproducible. This is in contrast to the high level of data shared in NLP (Mieskes, 2017), which should aid reproduction.

A way forward We need to consider the value as well as the cost of computation in terms of resources, people, and money. If it becomes impossible to replicate some experiments because of those factors, we need to foster computationally-affordable solutions that can be tested by everyone.

5 Unexplainable Methods

A final important issue is the low explainability of our models. This argument recalls the rationalism vs. empiricism debate we mentioned earlier, and is valid for most deep learning models. However, it has recently become more prevalent due to the low effort needed to set up and run these models.

GPT-3 (Brown et al., 2020) made it to the public quickly, most notably as "author" of an auto-generated Guardian article (GPT-3, 2020), which also generated a lot of opinions. The final article was the heavily-edited output of several GPT-3 runs to get a coherent version. So while still not an autonomous contribution, the potential impact that this technology is clear: It has allowed people to create eye-catching, interesting applications that capture the public's imagination. Those creations go beyond natural language. For example, GPT-3 can output HTML and CSS just from an input description of the desired outcome (Szőgyényi, 2020).

So while sensational results are picked up by the media and easily make their way to the general public (GPT-3, 2020), more nuanced comments and limitations of those results tend to be confined to in-domain newsfeeds (Marcus and Davis, 2020; Floridi and Chiriatti, 2020). Indeed, the public is left with the idea that these methods are either a panacea for all problems of the day, or the next step on the path to machine domination. But as scientists, we should realize that sensationalizing what we do comes with great responsibilities.

Easy availability of this technology can bring harm: it is not difficult to imagine the consequences of early access to this kind of technology, like bi-

ased auto-generated articles and fake news. However, it is difficult to predict which unwanted outcomes as of yet unknown models could generate. This is not a new argument: face recognition models have already created problems with racial bias, such as identifying images of Black people as "gorillas" (Zhang, 2015).

This point is even more poignant when we realize that different studies have pointed out that GPT-3 is not exhibiting what we as humans would call "intelligence" (Marcus and Davis, 2020; Floridi and Chiriatti, 2020) and have suggested that learning just from text and without including external experience might be a difficult task (Bisk et al., 2020; Bender and Koller, 2020; Bianchi et al., 2021a). Intelligent beings can explain their decisions and reason about their process. But GPT-3 and similar models make (unexplainable) decisions that *look* intelligent. This level of unexplainability hinders the future applications of this technology to areas that crucially depend on post-hoc explanations of the process. E.g., fields like medicine and law.

A way forward This problem does not have a solution – yet. We need to better engage with the media and public to make sure that what comes from our field is not only the great news of spectacular possibilities. It is difficult to also share the limitations: it might bore the public, detract from the undeniable successes, and it requires more effort to explain than simple, glowing success stories. But it is the only way to be sure that everyone has understood the full range of possible outcomes of unexplainable models.

6 Conclusion

We have argued that the current publication model of NLP fosters a gap between adoption and understanding of models, making it easier to meet publication demands with method papers, while shelving more unwieldy studies that include negative results and more epistemological approaches.

This issue is compounded by the rise of ever-larger models, which are unobtainable by all but a few researchers, and make it harder to explain how our methods work. As a result, reproducibility might suffer, and consequently endanger NLP's credibility and the public's trust in the field.

We do not make the argument that these are the only issues that our community has to take care of. For example, environmental sustainability (Strubell et al., 2019) of the models and the possible dual-use problem (Leins et al., 2020) are important topics that require a separate discussion.

Ethical Considerations

The main topic of this paper, reproducibility, is related to issues of ethics in NLP, with respect to fairness and accessibility. With this paper, we hope to contribute to that literature. Our paper does not contain new data sets or methods that pose a potential dual-use or bias problem.

References

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 381–393, Online. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. *Proceedings of FAccT*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Federico Bianchi, Ciro Greco, and Jacopo Tagliabue. 2021a. Language in a (search) box: Grounding language learning in real-world human-machine interaction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4409–4415, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021b. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021c. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational*

Linguistics: Main Volume, pages 1676–1683, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kevin Cohen, Karën Fort, Margot Mieskes, and Aurélie Névéol. 2020. Reviewing natural language processing research. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 16–18, Online. Association for Computational Linguistics.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251).

Caroline Criado Perez. 2019. *Invisible women: Exposing data bias in a world designed for men*. Random House.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, pages 101–109. ACM.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. 2020. A fair comparison of graph neural networks for graph classification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Antske Fokkens, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen, and Nuno Freire. 2013. Offspring from reproduction problems: What replication failure teaches us. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1691–1701, Sofia, Bulgaria. Association for Computational Linguistics.

GPT-3. 2020. A robot wrote this entire article. are you scared yet, human? — GPT-3. *The Guardian*.

John PA Ioannidis. 2005. Why most published research findings are false. *PLoS medicine*, 2(8):e124.

John PA Ioannidis. 2008. Why most discovered true associations are inflated. *Epidemiology*, pages 640–648.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988, Online. Association for Computational Linguistics.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep entity matching with pre-trained language models. *Proc. VLDB Endow.*, 14(1):50–60.

Gary Marcus and Ernest Davis. 2020. Gpt-3, bloviator: Openai's language generator has no idea what it's talking about. *Technology Review*.

Margot Mieskes. 2017. A quantitative study of data in the NLP community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29, Valencia, Spain. Association for Computational Linguistics.

Margot Mieskes, Karën Fort, Aurélie Névéol, Cyril Grouin, and Kevin Cohen. 2019. Community perspective on replicability in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 768–775, Varna, Bulgaria. INCOMA Ltd.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229. ACM.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Computer Vision – ECCV 2020*, pages 681–699, Cham. Springer International Publishing.

Peter Norvig. 2017. On chomsky and the two cultures of statistical learning. In *Berechenbarkeit der Welt?*, pages 61–83. Springer.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2020. What the [MASK]? making sense of language-specific BERT models. *arXiv preprint arXiv:2003.02912*.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. HONEST: Measuring hurtful sentence completion in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online. Association for Computational Linguistics.

Nils Reimers 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020a. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Anna Rogers, João Sedoc, and Anna Rumshisky, editors. 2020b. *Proceedings of the First Workshop on Insights from Negative Results in NLP*. Association for Computational Linguistics, Online.

Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. 2020. You CAN teach an old dog new tricks! on training knowledge graph embeddings. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Zoltán Szőgyényi. 2020. We built an OpenAI powered Tailwind CSS code generator using GPT-3. *Themesberg Magazine*.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270. Association for Computational Linguistics.

Yury Zemlyanskiy, Sudeep Gandhe, Ruining He, Bhargav Kanagal, Anirudh Ravula, Juraj Gottweis, Fei Sha, and Ilya Eckstein. 2021. DOCENT: Learning self-supervised entity representations from large document collections. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2540–2549, Online. Association for Computational Linguistics.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.

Maggie Zhang. 2015. Google photos tags two African-Americans as gorillas through facial recognition software. *Forbes Magazine*.