

PoliMi-FlatEarthers at CheckThat! 2022: GPT-3 applied to claim detection

Stefano Agresti¹, S. Amin Hashemian¹ and Mark J. Carman¹

¹Politecnico di Milano, Piazza Leonardo Da Vinci 32, Milan, Italy

Abstract

A crucial task when fighting online misinformation is the automatic filtering of potential claims among the millions of posts and texts uploaded every day to social media. Most approaches to this problem have focused on the training and fine-tuning of BERT-related models [1][2]. We show how much larger GPT-3 [3] models, despite being developed primarily for text-generation, outperformed previous language models on the task of automated claim detection on the 2022 CheckThat! Challenge [4] dataset. Not only that, we will also show that GPT-3, while designed for handling mainly English tasks, can maintain competitive performances on other languages as well.

Keywords

Automated claim detection, Automated fact-checking, GPT-3

1. Introduction

Over the last few years, online misinformation has been at the center of researchers' attention. Several methods have been proposed to counter the phenomenon, mainly implementing different forms of Artificial Intelligence and Natural Language Processing to filter the millions of posts and texts published daily on social media [5].

One of the solutions, outlined in [6], is the creation of an automated fact-checking system that replicates the fact-checking process of human fact-checkers through the use of AI. Researchers divided the problem into different steps: identifying claims worth fact-checking, detecting relevant previously fact-checked claims, retrieving relevant evidence to fact-check a claim, and actually verifying a claim. The first step, which consists in identifying, inside a piece of text, those sentences that constitute claims, is called claim-detection and is at the center of the first task of the CheckThat 2022! Challenge [4] [7] and of the CheckThat 2021! Challenge [2]. Most of the teams who took part in the first task in 2021, which was divided in 6 languages and 2 subtasks, employed automated text classifiers based on BERT-related models. The top-ranking teams across almost all of the languages and subtasks were using BERT [8] [9].

In 2020 OpenAI released GPT-3 [3], a new language model with a massive number of parameters (175 billion) trained for text generation. The autoregressive Language Model was shown to achieve strong performances on several NLP tasks, even with limited or no fine-tuning, using


CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ stefano.agresti19@gmail.com (S. Agresti); seyedamin.hashemian@mail.polimi.it (S. A. Hashemian); mark.carman@polimi.it (M. J. Carman)

🌐 <https://github.com/steflyx> (S. Agresti)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

instead zero-shot or few-shot learning.

In this paper, we show that GPT-3 based models can outperform BERT-based ones on the claim-detection task. We report the results we obtained over the 4 subtasks of task 1 of the CheckThat! 2022 challenge, showing that the models ranked in top positions in the English Claim Detection (first place) and Checkworthy Claim Detection (third place) subtasks. We also show that the model maintained competitive performances on 4 of the other languages available from the challenge (Arabic, Bulgarian, Spanish, Dutch) *without performing any further fine-tuning* on training data from those languages.

2. Related works

In the CheckThat 2021 challenge [2], different methods were proposed to tackle task 1: claim detection. Yet, almost all participating teams employed pre-trained BERT models, either multilingual or fine-tuned on the language specific to the subtask. Many of them resorted to Data Augmentation or Text Preprocessing to boost the performances of their models. Famous techniques used for Data Augmentation were lexical substitution, machine translation, weak supervision, and cross-lingual training [9] [10] [11]. To achieve better results, some teams leveraged an ensemble approach [12].

In July 2020, OpenAI released its massive GPT-3 language model. While designed for text-generation purposes, it was shown to bring, through fine-tuning, or even through simple one-shot or few-shot learning, significant improvements on several NLP benchmarks. Among researchers, as well as in the world of journalism ¹, the performances of GPT-3 raised concerns on whether it might be used maliciously to generate fake news [13], with several articles noting that it could be hard for human readers to know the differences between articles written by humans and GPT-3 [14] [3]. Others however, have shown that GPT-3 can be exploited to fight misinformation [15], applying its text classification abilities to discriminate between COVID-related tweets containing truthful and false information. The results obtained demonstrated that GPT-3 can be a useful asset in the fight against misinformation, leading us to decide to use it during our participation to the CheckThat! 2022 challenge.

3. Method

The GPT-3 model is so large that it is only available via a cloud-based API. Access to this API requires payment, so to constrain the costs of our experiment, we limited ourselves to develop four models, one for each of the subtasks for Task 1 in English, to stay below the threshold for free-tier usage. For task 1A, 1B and 1D we used the Curie GPT-3 model, currently the largest GPT-3 model available for fine-tuning², while for task 1C we experimented with the less powerful, but cheaper, Ada model. For each of the tasks, we used the following prompts and completions:

¹<https://www.wired.com/story/ai-write-disinformation-dupe-human-readers/>

²According to OpenAI, their largest model is DaVinci, but it is not available for fine-tuning

- **Subtask 1A:**
Prompt: “[Entry] Result.”;
Completion: “Check-Worthy” / “Not Check-Worthy”;
- **Subtask 1B:**
Prompt: “[Entry] Does it contain a verifiable claim?”;
Completion: “Yes” / “No”;
- **Subtask 1C:**
Prompt: “[Entry] Harmful.”;
Completion: “Yes” / “No”;
- **Subtask 1D:**
Prompt: “[Entry] Harmful.”;
Completion: “[Entry-Associated Label]”;

Where [Entry] was replaced with the training dataset entries. In all cases, we used the hyperparameters suggested by OpenAI:

- Number of training epochs: 4.
- Batch size: 0.2% of the size of the training dataset.
- Learning rate: 0.1.

Once the models were had been fine-tuned, we ran them on each of the entries from the testing datasets. Interestingly, the models reacted differently to the completions we passed them. The models for subtasks 1A and 1B always produced outputs starting with the completions we gave (i.e., *Check-Worthy/Not Check-Worthy* and *Yes/No*), so we just had to cut those outputs to obtain the proper classification format. Models for 1C and 1D tended instead to produce more varied outputs. In particular, for subtask 1D, which featured multiple classes, our models produced several outputs that had to be shortened to fit the class labels (an example was the model output: "yes contains advice for cure" which was shortened and assigned to the class: *yes_contains_advice*).

Although we only fine-tuned our models in English, we decided to test them on other languages to evaluate how GPT-3 would perform in this scenario. Again, to reduce our expenses, we didn’t perform this test for all subtasks, but only for subtask 1A in Arabic, Bulgarian, Dutch, and Spanish.

4. Results

The results of applying GPT-3 to these tasks are shown in Table 1 and were beyond our expectations, with our model outperforming all other models on subtask 1B (Accuracy of 0.761), while placing 3rd in subtask 1A (F1-positive of 0.626). This is particularly impressive if we consider that we didn’t perform any Text Preprocessing or Data Augmentation.

Results were mixed for subtask 1A on other languages, with low scores in Arabic and Bulgarian (F1-positive of 0.321 and 0.341, 5th position in both) and average scores in Dutch and Spanish (F1-positive of 0.532 and 0.323, position 4th and 2nd). However, considering that in these cases,

Table 1

Results of the GPT-3 models over the different subtasks

Subtask	Language	Score	Ranking
1A	English	F1-positive: 0.626	3/14
1A	Arabic	F1-positive: 0.321	5/5
1A	Bulgarian	F1-positive: 0.341	5/6
1A	Dutch	F1-positive: 0.532	4/6
1A	Spanish	F1-positive: 0.323	2/4
1B	English	Accuracy: 0.761	1/10
1C	English	F1-positive: 0.270	10/12
1D	English	F1-weighted: 0.636	7/7

not only did we not perform any Text Preprocessing or Data Augmentation, but we were also using models fine-tuned in English, these results can be considered more than satisfying.

Results for subtasks 1C and 1D were instead less positive (F1-positive of 0.270 for subtask 1C and F1 weighted of 0.636 for subtask 1D). As we noted in Section 3, these models were more likely to produce imprecise outputs (meaning that GPT-3 tended to produce text less likely to resemble the exact class labels defined for the problem). This could be a symptom indicating that GPT-3 was not able to identify the classes inside the training datasets correctly. It's worth noting that in subtask 1D (a multi-class classification task with 8 class labels), our model produced outputs that in some cases mixed some of the classes, for example the model output "yes contains advice for cure", which mixes the classes *yes_contains_advice* and *yes_discusses_cure*.

5. Conclusion

This paper showed that GPT-3 is competitive with BERT on claim-detection tasks, even without any form of Text Preprocessing or Data Augmentation. Our models remained competitive across different languages, despite being trained solely on English, showing their potential as multilingual classifiers.

Given larger resources, it would be useful to run more tests in the future, in order to clearly assess the effectiveness of GPT-3 in this kind of tasks. We point to the following ideas:

- Using the DaVinci model (the largest GPT-3 model, not available for fine-tuning) in a zero-shot, or few-shot, modality, instead of fine-tuning the less powerful models.
- Fine-tuning GPT-3 on a multi-lingual dataset and testing it on different languages.
- Fine-tuning GPT-3 on datasets from different languages, testing its performances on the same languages used for training.
- Applying Text Preprocessing and Data Augmentation techniques to the inputs, before fine-tuning and testing the models.

Although more research is needed to fully understand GPT-3 potentialities in the fight against fake news, given the results shown in this paper, and the flexibility demonstrated in our experiments, we believe that GPT-3 based models can indeed represent a valid asset for this task.

References

- [1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.
- [2] P. Nakov, G. Da San Martino, T. Elsayed, A. Barrón-Cedeño, R. Míguez, S. Shaar, F. Alam, F. Haouari, M. Hasanain, W. Mansour, B. Hamdan, Z. S. Ali, N. Babulkov, A. Nikolov, G. K. Shahi, J. M. Struß, T. Mandl, M. Kutlu, Y. S. Kartal, Overview of the clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer International Publishing, Cham, 2021, pp. 264–291.
- [3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- [4] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, J. M. Struß, T. Mandl, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, G. K. Shahi, H. Mubarak, A. Nikolov, N. Babulkov, Y. S. Kartal, J. Beltrán, The CLEF-2022 CheckThat! Lab on fighting the covid-19 infodemic and fake news detection, in: M. Hagen, S. Verberne, C. Macdonald, C. Seifert, K. Balog, K. Nørvgå, V. Setty (Eds.), *Advances in Information Retrieval*, Springer International Publishing, Cham, 2022, pp. 416–428.
- [5] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, *ACM Comput. Surv.* 53 (2020). URL: <https://doi.org/10.1145/3395046>. doi:10.1145/3395046.
- [6] P. Nakov, D. Corney, M. Hasanain, F. Alam, T. Elsayed, A. Barrón-Cedeño, P. Papotti, S. Shaar, G. Da San Martino, Automated fact-checking for assisting human fact-checkers, in: Z.-H. Zhou (Ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, International Joint Conferences on Artificial Intelligence Organization*, 2021, pp. 4551–4558. URL: <https://doi.org/10.24963/ijcai.2021/619>. doi:10.24963/ijcai.2021/619, survey Track.
- [7] P. Nakov, A. Barrón-Cedeño, G. Da San Martino, F. Alam, R. Míguez, T. Caselli, M. Kutlu, W. Zaghouni, C. Li, S. Shaar, H. Mubarak, A. Nikolov, Y. S. Kartal, J. Beltrán, Overview of the CLEF-2022 CheckThat! lab task 1 on identifying relevant claims in tweets, in: N. Faggioli, Guglielmo and Ferro, A. Hanbury, M. Potthast (Eds.), *Working Notes of CLEF 2022—Conference and Labs of the Evaluation Forum, CLEF '2022*, Bologna, Italy, 2022.
- [8] I. B. Schlicht, A. F. M. de Paula, P. Rosso, Upv at checkthat! 2021: Mitigating cultural

- differences for identifying multilingual check-worthy claims (2021). URL: <https://arxiv.org/abs/2109.09232>. doi:10.48550/ARXIV.2109.09232.
- [9] E. Williams, P. Rodrigues, S. Tran, Accenture at checkthat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation, 2021. URL: <https://arxiv.org/abs/2107.05684>. doi:10.48550/ARXIV.2107.05684.
- [10] X. Zhou, B. Wu, P. Fung, Fight for 4230 at checkthat! 2021: Domain-specific preprocessing and pretrained model for ranking claims by check-worthiness, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 681–692. URL: <http://ceur-ws.org/Vol-2936/paper-57.pdf>.
- [11] M. S. Zengin, Y. S. Kartal, M. Kutlu, Tobb etu at checkthat! 2021: Data engineering for detecting check-worthy claims., in: CLEF (Working Notes), 2021, pp. 670–680.
- [12] B. Carik, R. Yeniterzi, SU-NLP at checkthat! 2021: Check-worthiness of turkish tweets, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 476–483. URL: <http://ceur-ws.org/Vol-2936/paper-37.pdf>.
- [13] A. Tamkin, M. Brundage, J. Clark, D. Ganguli, Understanding the capabilities, limitations, and societal impact of large language models, 2021. URL: <https://arxiv.org/abs/2102.02503>. doi:10.48550/ARXIV.2102.02503.
- [14] B. Buchanan, A. Lohn, M. Musser, K. Sedova, Truth, Lies, and Automation: How Language Models Could Change Disinformation, 2021. doi:<https://doi.org/10.51593/2021CA003>.
- [15] R. Joseph, Infodemiology and infoveillance of covid19 using gpt-3 (2021).