# Natural Language Processing for Fact-Checking and Claim Assessment

A language model based approach

Othman El houfi
*MSc Data Science & Machine Learning*
CY Cergy Paris University, France
othmanelhoufi@gmail.com

Dimitris Kotzinos
*University Professor*
CY Cergy Paris University, France
dimitrios.kotzinos@cyu.fr

***Abstract*—As false information and fake news are propagating throughout the internet and social networks, the need of fact-checking operations becomes necessary in order to maintain a truthful digital environment where general information can be reliably exploited whether in politics, finance or other domains. The need of this online claim assessment comes from the fact that fake news and false information can have a big negative impact on politics, economy (2016 USA Elections) and public health (COVID-19).**
**A number of solutions have been proposed to deal with this problem and limit the spread of false information, both manual and automatic. Of course the manual approaches done on websites such as *PolitiFact.com*, *FactCheck.org* and *Snopes.com* don't construct a viable solution for the long term as the speed and scale of information propagation increase exponentially rendering this manual fact-checking operation where human fact-checkers can't scale up at the same rate limited and incapable of solving the problem.**
**Here, we present our contribution in this regard: an automated solution for fact-checking using Wikipedia as a source of truth and a state of the art language models used today for NLP tasks (BERT, RoBERTa, XLNet...) in order to classify a given claim as *Supports*, *Refutes* or *Not enough information (NEI)*.**

***Index Terms*—Natural Language Processing, Language Model, Wikipedia, Information retrieval, Text processing, Natural Language Inferencing, Fact-Checking, Document retrieval, Fake-news.**

## I. Introduction

From a social and psychological perspective, humans have been proven irrational and vulnerable when differentiating between truth and false news (typical accuracy ranges between 55% and 58%) [1], thus fake news obtain public trust relatively easier than truthful news because individuals tend to trust fake news after repeated exposure (*Validity effect*), or if it confirms their pre-existing beliefs (*Confirmation bias*), or simply due to the obligation of participating socially and proving a social identity (*Peer pressure*). The social sciences are still trying to comprehend the biological motivations that makes fake news more appealing to humans.

On the other hand, the growth of social media platforms resulted in a huge acceleration of news spreading whether true or false. As of Aug. 2017, 67% [1] of Americans get their news from social media. These platforms even give the user the right to share, forward, vote and participate to online discussions. All of this made the problem of fake news spreading more and more dangerous, our economies for example, are not robust to the spread of falsity, false rumors have affected stock prices and the motivations for large-scale investments, as we witnessed after a false tweet claimed that Barack Obama was injured in an explosion which caused $130 billion drop in stock value [2]. Another recent example is related to public health where rumors about COVID-19 vaccines and drug companies influenced people in their decision on getting vaccinated.

That being said, is there a way to monitor the spread of fake news through social media? Or more specifically, how can we differentiate between fake news and truthful news, and at what level of confidence can we do that?

From a computer engineering perspective, different approaches were studied:

- **Knowledge-based Fake News Detection [3]:** a method aims to assess news authenticity by comparing the knowledge extracted from to-be verified news content with known facts, also called fact-checking.
- **Style-based Fake News Detection [4]:** focuses on the style of writing, i.e. the form of text rather than its meaning.
- **Propagation-based Fake News Detection [5]:** a principled way to characterize and understand hierarchical propagation network features. We perform a statistical comparative analysis over these features, including micro-level and macro-level, of fake news and true news.
- **Credibility-based Fake News Detection [6]:** the information about authors of news articles can indicate news credibility and help detect fake news.

In this paper we will focus on a new approach that utilizes Language Models (LMs) for fact-checking. The goal is not to implement an algorithm that scans social networks for real

time fake news detection, but rather we will create a model that can assess with a degree of confidence the truthfulness or falseness of a claim given by a user as an input by exploiting LMs that were already trained on Wikipedia, and fine-tune each LM for a downstream task in order to solve this classification problem.

## II. Related works

### A. Language model based approach [8] [9]

A paper called *"Language Models as Fact Checkers?"* done by a team from *FacebookAI* and *Hong Kong University of Science and Technology*, provides an example of a fact-checking model using zero-shot LM that outperforms a random baseline LM using the FEVER dataset [10].

The goal of fact-checking as mentioned previously, and relatively to this paper, is to validate the truthfulness of a given claim. Each claim is assigned to one of these labels: *Supports*, *Refutes* or *Not enough information (NEI)* to verify.

This paper describes the difference between Traditional Pipeline fact-checking models and their zero-shot fact-checking LM:

- **Traditional pipeline:** this type of models access knowledge within an external knowledge base like Wikipedia in order to validate a claim. It involves information retrieval modules such as document retrieval and sentence retrieval.
- **Zero-shot LM pipeline:** it replaces both the external knowledge base and the information retrieval modules with a pre-trained language model.

They used the publicly available 24-layer BERT-Large as our language model, which was pre-trained on Wikipedia in 2018. After fine-tuning the model they achieved 57% in accuracy and 57% in F1-macro score which was better than the baseline BERT model (without fine-tuning) that achieved 49% in accuracy and 44% in F1-macro score.

### B. Perplexity based approach [11] [12]

In March 2021, *Nayeon Lee*, *Yejin Bang*, *Andrea Madotto*, *Madian Khabsa*, and *Pascale Fung* published a paper called *Towards Few-Shot Fact-Checking via Perplexity* where they propose a new approach of the powerful transfer learning ability of a language model via a perplexity score. Using a method called *few-shot learning*, they built a model that outperforms major class baseline models by more than 10% on the F1-Macro metric score.
In this paper the goal is to determine the veracity of a claim given some evidence, for this they define a claim, evidence pair. The label *Supported* is assigned when relevant evidence exists that supports the claim, and *Unsupported* label for the opposite case.
*Unsupported* claims on average have higher perplexity than *Supported* claims. For example, *Supported* claim "Washing hands prevents the spread of diseases" has a perplexity value of 96.74, whereas the *Unsupported* claim "All dogs speak English fluently" has a much higher perplexity value of 328.23. The datasets used in this experiment are: Covid19-Scientific, Covid19-Social, and FEVER. As for the perplexity based experiment they used one unidirectional LM and one masked LM:

- $PPL_{GPT2-B}$ : a single-parameter classifier based on perplexity from GPT2-base [13] (unidirectional LM)
- $PPL_{BERT-B}$ : a single-parameter classifier based on perplexity from BERT-base [14] (Masked LM)

They took into consideration the accuracy and the F1-Macro metrics for the evaluation. Because the datasets are unbalanced, they mainly consider the F1-Macro score over accuracy as an overall evaluation. The perplexity-based classifiers, especially $PPL_{GPT2-B}$, outperform all Major Class baselines across all tasks in all settings. For instance, $PPL_{GPT2-B}$ achieved accuracy of 67.48% and F1-macro score of 64.70% on FEVER dataset.
On the other hand the classification was limited to two labels (*Supported* and *Unsupported*) which does not solve the entire classification problem in the FEVER dataset that provides three labels (*Supports*, *Refutes* or *Not enough information (NEI)*).

## III. Presented method

Most of the fact-checking algorithms today involving knowledge-based verification uses a traditional pipeline that puts in place a module for retrieving articles from an external source, another module for retrieving relevant sentences from each article, and a last module for natural language inferencing (NLI) in order to classify a claim.

In this paper we present a method that is fully reliant on the powerfulness of today's best LMs. We start by fine-tuning each model for the downstream task that is claim classification using the FEVER dataset, then each model is employed to assess the validity of new input claims. This approach takes into consideration only an internal knowledge source (FEVER) for fine-tuning, that is for the learning phase, which makes the prediction phase knowledge-free rather than utilizing external knowledge sources for retrieving articles and sentences.

It is also important to mention that we only use LMs for classifying claims and not for generating evidence. We leave generating evidences with language models for future work.

### A. Dataset

FEVER (Fact Extraction and VERification) consists of 185,445 claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. The claims are classified as Supported, Refuted or NotEnoughInfo. For the first two classes, the annotators also recorded the sentence(s) forming the necessary evidence for their judgment [10].

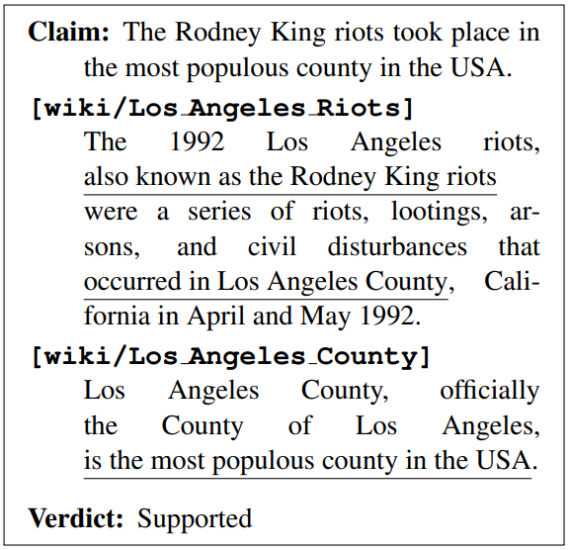


Fig. 1. Manually verified claim requiring evidence from multiple Wikipedia pages.

## IV. Results and Analysis

### A. Experimental protocol

### B. Results

### C. Complexity analysis

### D. Conclusion of the analysis

## V. Conclusion

## References


[1] Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 836–837, 2019.

[2] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[3] Anton Chernyavskiy, Dmitry Ilvovsky, and Preslav Nakov. Whatthewikifact: Fact-checking claims against wikipedia. *arXiv preprint arXiv:2105.00826*, 2021.

[4] Piotr Przybyla. Capturing the style of fake news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 490–497, 2020.

[5] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 626–637, 2020.

[6] Niraj Sitaula, Chilukuri K Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. Credibility-based fake news detection. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 163–182. Springer, 2020.

[7] Mykola Trokhymovych and Diego Saez Trumper. Natural language inference for fact-checking in wikipedia.

[8] Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. Language models as fact checkers? *arXiv preprint arXiv:2006.04102*, 2020.

[9] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[10] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

[11] Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung. Towards few-shot fact-checking via perplexity. *arXiv preprint arXiv:2103.09535*, 2021.

[12] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. Misinformation has high perplexity. *arXiv preprint arXiv:2006.04666*, 2020.

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.