Towards Few-Shot Fact-Checking via Perplexity

Nayeon Lee[†*] **Yejin Bang**[†*]
Andrea Madotto[†] **Madian Khabsa**[§] **Pascale Fung**[†]
[†]Hong Kong University of Science and Technology [§]Facebook AI
{nyleeaa,yjbang}@connect.ust.hk

Abstract

Few-shot learning has drawn researchers' attention to overcome the problem of data scarcity. Recently, large pre-trained language models have shown great performance in few-shot learning for various downstream tasks, such as question answering and machine translation. Nevertheless, little exploration has been made to achieve few-shot learning for the fact-checking task. However, fact-checking is an important problem, especially when the amount of information online is growing exponentially every day. In this paper, we propose a new way of utilizing the powerful transfer learning ability of a language model via a perplexity score. The most notable strength of our methodology lies in its capability in *few-shot* learning. With only two training samples, our methodology can already outperform the Major Class baseline by more than an absolute 10% on the F1-Macro metric across multiple datasets. Through experiments, we empirically verify the plausibility of the rather surprising usage of the perplexity score in the context of fact-checking and highlight the strength of our few-shot methodology by comparing it to strong fine-tuning-based baseline models. Moreover, we construct and publicly release two new fact-checking datasets related to COVID-19.

1 Introduction

Few-shot learning is being actively explored to overcome the heavy dependence on large-scale labeled data that serves as a crucial bottleneck to machine learning models. Recently, researchers have explored few-shot learning that leverages the powerful transfer learning ability of pre-trained large language models (LMs) in various NLP tasks. Petroni et al. demonstrated that an LM serves as a good zero-shot learner on the question-answering task due to its encoded commonsense knowledge.

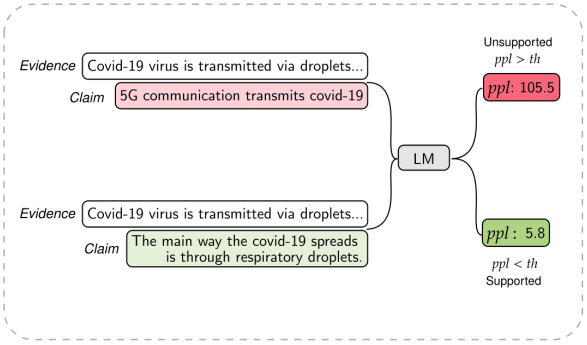


Figure 1: Illustration of our simple yet effective perplexity-based approach. Few-shot data samples are used to find the optimal perplexity threshold th that separates `Unsupported` claims from `Supported` claims.

Going further, Brown et al. illustrated the impressive potential of LMs as strong zero-shot and few-shot learners across translation, commonsense reasoning and natural language inference (NLI). However, little or no exploration has been made on few-shot learning in the fact-checking domain, which is a timely and important task in which data-scarcity is particularly problematic.

Previous works have proposed different ways of leveraging LMs to conduct zero- or few-shot learning. One common approach is to query the LM for the missing token (i.e., "answer") for the zero-shot question-answering task (Petroni et al., 2019; Brown et al., 2020) by transforming questions into a form of statement. Another approach is to adopt an in-context learning approach where the input context of the LM is carefully crafted to control the output. For example, a natural language task instruction (e.g., "Translate English to French:") or training sample (e.g., "sea otter => loutre de mer") is provided as the context for zero-shot/few shot translation (Brown et al., 2020).

In this work, we explore a new way of leveraging LMs for few-shot learning in the fact-checking task. This is done by leveraging a perplexity score from

[*] Equal contribution.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1971–1981
June 6–11, 2021.

Unsupported Claims	Perplexity	**Supported** Claims	Perplexity
5G network can spread diseases.	826.70	Beyonce is one of the most famous singers in the world.	23.03
All dogs speak English fluently.	328.23	Chicago is one of the cities in the United States.	43.92
Washing hands **helps** the spread of diseases.	201.10	Washing hands **prevents** the spread of diseases.	96.74

Table 1: Relations between veracity of claim and perplexity. `Unsupported` claims have higher perplexity compared to `Supported` claims. Note that the perplexity score listed here is using GPT2-base on each of the claims.

evidence-conditioned LMs. Fact-checking is the task of verifying a claim based on its corresponding evidence, and one of its most important objectives is to correctly model the relationship between the given claim and evidence. We hypothesize that a perplexity score from evidence-conditioned LMs is helpful for such purpose since perplexity measures the likelihood of a given sentence with reference to previously encountered text (i.e., given the evidence prefix and the LM's training corpus). Therefore, this paper attempts to investigate this hypothesis and proposes a novel perplexity-based few-shot learning methodology for fact-checking.

Through experimental analysis, we empirically demonstrate the effectiveness of our proposed methodology in *few-shot learning*, and we compare it to strong fine-tuning-based baselines. Moreover, we compare different LMs (BERT and GPT2) in different sizes, from small to XL, to unveil interesting insights on which model is more suitable for this task. Finally, we discuss the potential application of evidence-conditioned perplexity for ranking candidate claims in priority order of the most urgent to be fact-checked to the least.

Our contribution is three-fold: First, we propose an effective way of leveraging the perplexity score in the context of fact-checking. We would like to emphasize that our approach is a simple yet effective way of leveraging large pre-trained LMs. Second, we demonstrate the effectiveness of the perplexity-based approach in the *few-shot* setting by outperforming strong fine-tuned baselines, such as BERT (Devlin et al., 2019), RoBERTA (Liu et al., 2019), and XLNet (Yang et al., 2019), by an absolute $10 \sim 20\%$ F1-Macro scores in the 2-,10-, and 50-shot settings. Third, we construct two new fact-checking datasets related to COVID-19, which has caused the problem of an "infodemic".

2 Related Work

Fact-checking is a complex task that is split into many sub-tasks. First, credible sources of evidence need to be identified. Second, a set of relevant evidence needs to be retrieved from the identified credible sources. Last, veracity classification of claims can be made based on the retrieved evidence.

Some works have focused on full-pipeline systems that handle all sub-tasks and provide real working web prototypes (Karadzhov et al., 2017; Popat et al., 2017, 2018a; Hasanain et al., 2019; Tokala et al., 2019). These works use the entire Web as a knowledge source to confirm or reject a claim taking the credibility or reliability of the Web source into account. Another common setting for fact-checking is to assume a credible evidence source is given (e.g., Wikipedia), and to focus on the evidence retrieval and veracity verification steps only. FEVER (Thorne et al., 2018) and Tabfact (Chen et al., 2019) are two large datasets for this setting, and there are many follow-up studies working on them (Yoneda et al., 2018a; Nie et al., 2019; Zhong et al., 2020; Herzig et al., 2020; Zhou et al., 2019; Hidey et al., 2020).

Our work follows the latter group of works and uses the following setting: given a tuple consisting of claims and relevant evidence, we classify the final fact-checking veracity label of the given claim (Popat et al., 2018b; Ma et al., 2019; Wu et al., 2020). By doing this, we focus on the methodology for the veracity classification task without worrying about the propagated errors from earlier modules, such as source credibility profiling and evidence retrieval.

Leveraging LMs as a knowledge base, zero-shot learner or a few-shot learner has been gaining popularity within the NLP field. It was discovered that large pre-trained LMs can store factual knowledge in their parameters (Petroni et al., 2019; Roberts et al., 2020; Madotto et al., 2020), and that this stored knowledge can help LM to be good at zero-shot and few-shot learning in various NLP tasks, such as question answering, summarization, textual entailment, translation and commonsense reasoning (Brown et al., 2020). For the task of fact-checking, Lewis et al. and Lee et al. attempted to

leverage such LMs. However, they mainly use the model to replace the evidence retriever of the fact-checking pipeline, and they still require training of final veracity classifier. Our work, in contrast, focuses on the few-shot ability of LMs for *veracity classification*.

3 Preliminary Exploration of Hypothesis

In this section, we conduct a preliminary investigation to validate the potential of our hypothesis that the perplexity score from an evidence-conditioned LM can provide a signal for claims unsupported by evidence.

For our exploration, we first collect a small set of `Supported` and `Unsupported` claims that can be verified based on the training corpus of the target LM (namely, Wikipedia which is used in the training of many pre-trained LMs). Then, we compare the perplexity scores between them.

To recap, perplexity is a commonly used metric for measuring the performance of LMs. It is defined as the inverse of the probability of the test set normalized by the number of words:

$$PPL(X) = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{p(x_i|x_0, \ldots, x_{i-1})}}. \quad (1)$$

Another way of interpreting perplexity is as a measure of the likelihood of a given test sentence with reference to the training corpus.

From Table 1, we can observe that `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. We believe these observations support our hypothesis. Thus, we proceed to build our approach based on this hypothesis (Section 4), and conduct experiments (Section 5) and analysis (Section 6) to verify the validity of our perplexity-based fact-checking approach.

4 Methodology

4.1 Task definition

In this work, we define our task to be: Given a {claim, evidence} pair, determine the veracity of a claim against the evidence - i.e., `Supported` vs. `Unsupported` claims. The label `Supported` is assigned when there exists relevant evidence that supports the claim, while `Unsupported` is assigned when there does not exist any supporting evidence. Note that this existence of refuting evidence also places a claim into this latter category.

4.2 Evidence Conditioned Perplexity

Although previous works have shown that an LM can encode knowledge from its training corpus, there are a few limitations to solely relying on the pre-trained weights. First, we cannot easily check and guarantee whether the LM has already seen the evidence that is required for verification, and the LM would definitely not have seen the evidence related to newly emerging events after the LM pre-training. For instance, the event of COVID-19 emerged after the release of the GPT2 pre-trained model. Second, although LMs have shown surprising ability in memorizing some knowledge, they are not perfect, as pointed out by previous works (Poerner et al., 2019; Lee et al., 2020). Therefore, we propose to incorporate evidence into the perplexity calculation by using it as a prefix of the claim.

There are two popular kinds of LMs: i) unidirectional LMs that are trained with the conventional next token prediction task, and ii) masked LMs that are trained with the masked token prediction token, resulting in a bidirectional LM. We briefly describe how to obtain the evidence-conditioned perplexity for both types of LM:

Unidirectional Language Model Perplexity For a unidirectional LM, first we concatenate the evidence and claim to obtain the input to the LM: $X = \{x_{e_0}, \ldots, x_{e_E}, x_{c_0} \ldots, x_{c_C}\}$, where E and C denote the number of evidence tokens and claim tokens, respectively. Then, we obtain the evidence-conditioned perplexity by

$$PPL(X) = \sqrt[C]{\prod_{i=1}^{C} \frac{1}{p(x_{c_i}|x_{e_0}, \ldots, x_{e_E}, \ldots, x_{c_{i-1}})}}.$$

Note that the evidence tokens are used to condition the perplexity, yet their conditional probabilities $p(x_{e_i}|x_{e_0}, \ldots, x_{e_{i-1}})$ do not contribute to the $PPL(X)$, which is the main difference from Eq. (1).

Data sets	Unsupported claims	Supported claims	Total
Covid19-Scientific	101	71	172
Covid19-Social	263	77	340
FEVER	3333	3333	6666

Table 2: Dataset Statistics

Masked Language Model Pseudo Perplexity A masked LM (MLM) is a type of LM, first proposed by Devlin et al., which is trained with the masked token prediction task instead of the next token prediction task. The "perplexity" score from the MLM does not mean the same as the conventional perplexity score. Therefore, we use the "pseudo perplexity" score proposed by Salazar et al., which is computed by summing all the log probabilities obtained by sequentially masking each token in the input sentence.

4.3 Leveraging Perplexity

Once we obtain the evidence-conditioned perplexity scores for each claim, we find the best threshold th that separates `Supported` claims from `Unsupported` claims. We would like to emphasize that our approach does not involve any parameter update of the LM. We only do inference with the LM, and leverage the few-shot samples as the "validation set" to find the optimal single threshold parameter, th. Throughout our paper, we refer to our methodology as the "perplexity-based classifier".

Given a set of a claim and evidence, if the evidence-conditioned perplexity score is less than the threshold (i.e. $< th$), the claim is `Supported` by the evidence; otherwise it is `Unsupported`.

5 Few-shot Experiment

5.1 Dataset[1]

All datasets used in the experiment are in English, and we report the data statistics in Table 2.

Covid19-Scientific A new test set is constructed by collecting COVID-19-related myths and scientific truths labelled by reliable sources like MedicalNewsToday, the Centers for Disease Control and Prevention (CDC), and the World Health Organization (WHO). It consists of the most common scientific or medical myths about COVID-19, which must be debunked correctly to ensure the safety of the public (e.g., "drinking a bleach solution will prevent you from getting COVID-19"). The set contains 172 claims with labels (`Supported`, `Unsupported`) obtained from the aforementioned reliable sources. Note that myths that are unverifiable from current findings are also assigned the `Unsupported` label.[2]

The gold evidence is obtained from the winning system of the Kaggle Covid-19 challenge (Su et al., 2020). This system retrieves the evidence from 59,000 scholarly articles about COVID-19, SARS-CoV-2, and other related corona viruses.[3]

Covid19-Social Another test set is constructed by crawling 340 COVID-19-related claims fact-checked by journalists from a website called Politifact.com. Unlike the Covid19-Scientific dataset, it contains non-scientific and socially-related claims, such as "For the coronavirus, the death rate in Texas, per capita of 29 million people, we're one of the lowest in the country." Such claims may not be life-and-death matters, but they still have the potential to bring negative sociopolitical effects. Originally, these claims are labelled into six classes {pants-fire, false, barely-true, half-true, mostly-true, true}. However, we use it in a binary setup for consistency with the Covid19-Scientific setup by assigning the first three classes to `Unsupported` and the rest to `Supported`.

For evidence of each claim, we follow the Alhindi et al. to obtain the human-written evidence/justification available on the Politifact.com website, from which the claims are crawled.

FEVER (Thorne et al., 2018) Fact Extraction and Verification (FEVER) is a publicly released large-scale dataset generated by altering sentences extracted from Wikipedia to promote research on fact-checking systems. Since our few-shot experiment requires little data, we only leverage the "Paper Test Dataset" from the FEVER workshop (https://fever.ai/) resource page to speed up our experiments.

This dataset originally has three classes, {Support, Refute, Not Enough Info}. "Support" is sim-

[1]Authors from HKUST obtained performed all experiments with the existing datasets and compiled and released the new datasets.

[2]Disclaimer: The data were collected during the early outbreak of COVID-19 (March 2020). The veracity may have been updated as the time evolved, but we release the original version of the dataset for future comparison

[3]https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

ilar to our Supported label, where a claim can be supported by given evidence. “Refute" is where a claim is "refuted" by given evidence, whereas “Not Enough Info" means not enough evidence is available for verification. For our FEVER experiment, we treat “Refute” and “Not Enough Info” as one class. This is because we believe that in a real scenario both cases are Unsupported claims that need attention.

To provide further detail, the “Support" class is mapped into Supported, and “Refute”/“Not Enough Info” is mapped into Unsupported to match our task setting. Note that to balance the dataset, we obtain half the data from “Refute” and the other half from “Not Enough Info”. Note that the gold evidence is included in the dataset released by Thorne et al.

5.2 Models

Ours We consider one unidirectional LM and one masked LM for our proposed perplexity-based methodology.

- $\texttt{PPL}_{\texttt{GPT2-B}}$ – Our single-parameter classifier based on perplexity from GPT2-base (Radford et al., 2019) (unidirectional LM)
- $\texttt{PPL}_{\texttt{BERT-B}}$ – Our single-parameter classifier based on perplexity from BERT-base (Devlin et al., 2019) (Masked LM)

Baselines We *finetune* various pre-trained Transformer-based (Vaswani et al., 2017) models to build our baseline classifiers, which is a common approach used to achieve many state-of-the-art results in the literature.

- Major Class – A simple majority classifier which always assigns the majority class of the training set to all samples. We provide this for reference because some of our dataset classes are imbalanced.
- $\texttt{BERT-B}_{\texttt{ft}}$ – A fine-tuned BERT-base model with a feed-forward classifier trained on top.
- $\texttt{BERT-L}_{\texttt{ft}}$ – A fine-tuned BERT-large model with a feed-forward classifier trained on top.
- $\texttt{RoBERTa}_{\texttt{ft}}$ – A fine-tuned RoBERTa-base model (Liu et al., 2019) with a feed-forward classifier trained on top.
- $\texttt{XLNet}_{\texttt{ft}}$ – A fine-tuned XLNet-base model (Yang et al., 2019) with a feed-forward classifier trained on top.

5.3 Experimental Setup

Few-Shot Data Setup Given N_D as the size of the dataset D, we do an n-shot experiment with n samples from D as a “validation set” for our perplexity-based approach or as a “training set” for the fine-tuning approach, and the remainder $(N_D - n)$ as a test set. To give a concrete example, in the 2-shot experiment using the Covid19-Social dataset (340 samples), we have two training samples and 338 test samples. We use three seeds to split the datasets and train the models. For a fair comparison, all the seeds and splits are kept the same across the models.

Evaluation We mainly evaluate our experiments using accuracy and the Macro-F1 metric. Since some of our datasets are imbalanced (the ratio of Supported to Unsupported in Table 2), we prioritize the overall Macro-F1 score over accuracy.

Training Details In our methodology, no gradient update is required. Thus, there are no training details such as learning rate, batch size or max-epoch to report. We simply use a small validation set (size of 2,10,50) to find the best-performing hyper-parameter value for the threshold th from the range of $\{0 \sim 1000\}$. None of the samples from the test set were seen in threshold searching.

For baseline fine-tuned classifiers, we do a grid-search to find the best-performing parameters, as follows: We use a learning rate of $5\text{e}{-6}$ for training the $\texttt{BERT-B}_{\texttt{ft}}$, $\texttt{RoBERTa}_{\texttt{ft}}$, and $\texttt{XLNet}_{\texttt{ft}}$ models, while $\texttt{BERT-L}_{\texttt{ft}}$is trained with a rate of $2\text{e}{-5}$. All models share the same batch size of 32 and maximum input sequence length of 128. We also use early-stopping with patience 3 with a maximum of 10 training epochs. Each experiment is run on an Nvidia GTX 1080 Ti, and each epoch takes $2 \sim 15$ seconds depending on the number of the training samples n. Note that for reproducibility, we will also publicly release the code.

5.4 Experimental Results

Table 3 reports the few-shot performance of the fine-tuning-based baselines and our perplexity-based classifiers.

Usage of Perplexity We can observe that our perplexity-based classifiers, especially $\texttt{PPL}_{\texttt{GPT2-B}}$, outperform all Major Class baselines across all tasks in all settings. For instance, $\texttt{PPL}_{\texttt{GPT2-B}}$ outperforms the Major Class by a great margin of 16% and 36.8% on accuracy and F1-Macro scores,

Shot #	Models	Fine-tuning?	Size	Covid-Scientific		Covid-Social		FEVER	
				Acc	F1-Macro	Acc	F1-Macro	Acc	F1-Macro
	Major Class	N/A	N/A	58.72%	37.00%	77.35%	43.62%	50.00%	33.33%
2	BERT-B_{ft}	yes	110M	47.34%	32.21%	26.11%	23.33%	51.56%	37.34%
	BERT-L_{ft}	yes	336M	49.39%	34.80%	37.78%	27.81%	50.80%	36.49%
	RoBERTa$_{ft}$	yes	125M	52.66%	34.29%	40.75%	26.78%	50.00%	33.33%
	XLNet$_{ft}$	yes	110M	51.48%	48.49%	57.67%	44.35%	49.41%	44.65%
	PPL_{GPT2-B}	no	117M	**66.75%**	**64.39%**	62.61%	**53.61%**	**61.92%**	**57.50%**
	PPL_{BERT-B}	no	110M	47.93%	38.54%	**77.74%**	49.15%	52.54%	41.33%
10	BERT-B_{ft}	yes	110M	46.27%	31.70%	43.26%	30.70%	51.56%	37.34%
	BERT-L_{ft}	yes	336M	50.00%	36.74%	60.49%	42.18%	50.80%	36.49%
	RoBERTa$_{ft}$	yes	125M	52.64%	40.28%	40.73%	26.73%	50.00%	33.33%
	XLNet$_{ft}$	yes	110M	49.69%	42.44%	59.68%	39.45%	49.41%	44.65%
	PPL_{GPT2-B}	no	117M	**72.98%**	**68.57%**	**71.23%**	**55.11%**	**62.82%**	57.04%
	PPL_{BERT-B}	no	110M	63.15%	60.77%	61.90%	46.35%	57.59%	**57.11%**
50	BERT-B_{ft}	yes	110M	56.75%	53.61%	60.21%	36.91%	52.18%	38.82%
	BERT-L_{ft}	yes	336M	56.75%	39.15%	64.94%	44.07%	51.14%	39.99%
	RoBERTa$_{ft}$	yes	125M	56.40%	38.97%	73.13%	45.30%	50.44%	38.15%
	XLNet$_{ft}$	yes	110M	63.22%	51.98%	**77.62%**	43.70%	49.18%	48.42%
	PPL_{GPT2-B}	no	117M	**74.73%**	**73.83%**	73.63%	**59.91%**	**67.48%**	**64.70%**
	PPL_{BERT-B}	no	110M	62.53%	61.11%	71.11%	54.72%	57.44%	56.94%

Table 3: Results comparison among perplexity-based classifiers and fine-tuned classifiers in 2-shot, 5-shot and 10-shot settings across three different tasks. Models whose names start with PPL are our proposed perplexity-based classifiers. Major Class is a reference to evaluate classifier performance. All test results reported are mean values of three trials with randomly selected n-shot training samples from the dataset, where $n = \{2, 10, 50\}$.

for the Covid-Scientific dataset in the 50-shot setting. This supports our hypothesis that evidence-conditioned perplexity scores are capable of providing signals regarding the veracity of the given claim.

Intuitively, we can consider the perplexity score to be mimicking the role of the "logits" from a classifier, and we are trying to find the best threshold to map this pseudo-logit-like perplexity score into a veracity label. The classification performance of our perplexity-based approach increases as the shot size increases. As the shot size increases from 2 to 50, PPL_{GPT2-B} shows an average gain of $8.19 \pm 2.74\%$ and $7.64 \pm 1.61\%$ in accuracy and Macro-F1 score, respectively, across all tasks. This is because a greater number of data samples means more anchor perplexity points for threshold searching, and thus, a better threshold to determine the veracity of claims.

Few-shot Comparison to Fine-tuned Baselines Except for the Covid-Social accuracy in the 50-shot setting, both of our proposed classifiers (PPL_{GPT2-B} , PPL_{BERT-B}) outperform the fine-tuned baseline classifiers across all tasks in all of the 2-, 10- and 50-shot settings. For the 2-shot and 10-shot settings, many of the baseline classifiers underperform the Major Class baseline regardless of the task. This implies their failure to learn anything from the fine-tuning step with a limited number of samples. Only after 50-shot do these baselines start to learn and outperform the Major Class baselines. This is not surprising, since the pre-trained models are known to perform well in a full-shot scenario, but they do not guarantee good performance when they are shown few samples.

In contrast, our perplexity-based classifiers manage to perform fairly well, even in the 2-shot setting, because our "classifier" is a single parameter (i.e., threshold value), which requires no complex learning or optimization. We would like to emphasize that ours consistently outperform the strong Transformer-based baselines across all dataset on the F1-Macro metric by absolute $10 \sim 20\%$. We argue that these results demonstrate the strength of our approach in low-resource few-shot settings.

BERT vs. GPT2 for Perplexity Scores Most of the time, PPL_{GPT2-B} outperforms PPL_{BERT-B}. For instance, in the 50-shot setting for the FEVER dataset, performance differences are 10.04% and 7.76% for accuracy and F1-Macro

LM Type	Parameter Size	Covid-Scientific		Covid-Social		FEVER	
		Acc	F1 Macro	Acc	F1 Macro	Acc	F1 Macro
$\texttt{PPL}_{\texttt{GPT2-B}}$	117M	74.73%	73.83%	73.63%	59.91%	67.48%	64.70%
$\texttt{PPL}_{\texttt{GPT2-M}}$	345M	75.11%	73.93%	**75.43%**	**60.23%**	69.02%	66.39%
$\texttt{PPL}_{\texttt{GPT2-L}}$	774M	76.19%	75.53%	73.29%	59.30%	71.66%	69.99%
$\texttt{PPL}_{\texttt{GPT2-XL}}$	1558M	**78.23%**	**77.63%**	72.80%	59.88%	**73.67%**	**71.71%**

Table 4: Effect of LM parameter size on the performance of proposed perplexity-based approach in 50-shot setting. All the results are the mean value of three trials.

Shot #	Ablation	Covid-Scientific		Covid-Social		FEVER	
		Acc	F1 Macro	Acc	F1 Macro	Acc	F1 Macro
2	$\texttt{PPL}_{\texttt{GPT2-XL}}$	**68.52%**	**66.21%**	**66.62%**	**52.68%**	**62.37%**	**56.35%**
	− evidence-conditioning	62.92%	59.53%	64.32%	51.37%	54.72%	45.65%
50	$\texttt{PPL}_{\texttt{GPT2-XL}}$	**78.23%**	**77.63%**	**72.80%**	**59.88%**	**73.67%**	**71.71%**
	− evidence-conditioning	73.35%	70.21%	71.97%	56.08%	56.69%	47.58%

Table 5: Ablation study – Effect of the evidence-conditioning on the classification performance.

scores respectively. Based on this observation, we can speculate that the perplexity from a unidirectional LM is more suitable for our proposed method than from a masked LM. This is most likely because the BERT perplexity score is only an estimation based on the "pseudo-perplexity" proposed by Salazar et al.

6 Analysis and Discussion

In this section, we conduct multiple analysis to further evaluate and understand aspects of our perplexity-based approach.

6.1 Scaling the Language Model Size

Generally, scaling the model size helps to also improve the model performance, because more parameters mean a stronger learning capability during fine-tuning or training. Also, Roberts et al. have demonstrated that increasing the parameter size allows for more knowledge to be packed into the LM's parameters. Therefore, we experiment with the model size to see if such findings also extend to our proposed methodology. The following model sizes of GPT2 are investigated: base ($\texttt{PPL}_{\texttt{GPT2-B}}$), medium ($\texttt{PPL}_{\texttt{GPT2-M}}$), large ($\texttt{PPL}_{\texttt{GPT2-L}}$) and xl ($\texttt{PPL}_{\texttt{GPT2-XL}}$).

Results are reported in Table 4. As expected, we can observe the trend that the performance increases with parameter size. For instance, $\texttt{PPL}_{\texttt{GPT2-XL}}$ is the best performing compared to the other, smaller, models for Covid-Scientific and FEVER, achieving the new state-of-the-art few-shot results by gaining absolute $\sim 4\%$ on Covid-Scientific and $\sim 2\%$ on FEVERfor accuracy/F1-Macro.

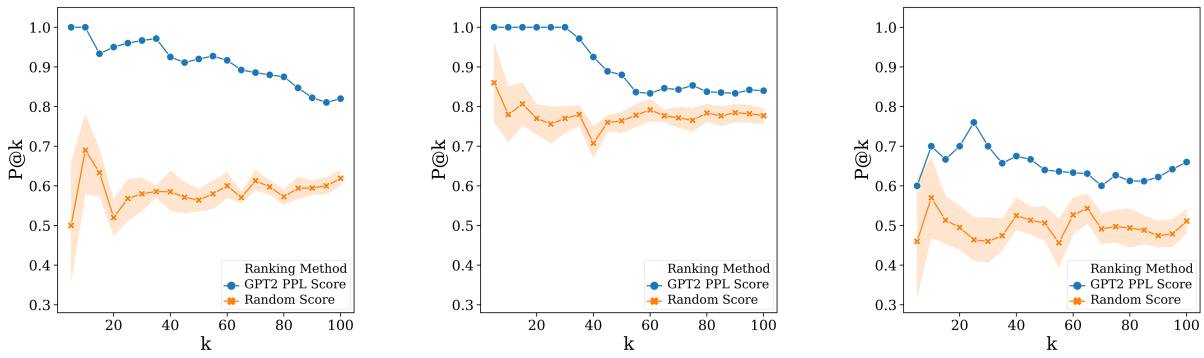
6.2 Ablation Study

We carry out an ablation study on the effect of evidence-conditioning in respect of the final perplexity scores and the corresponding final classification performance. In Table 5, we can observe the performance drops when evidence-conditioning is ablated – the biggest drop is $\sim 15\%$ on F1-Macro for the FEVER task in the 50-shot setting. This implies that the perplexity score is assigned *in relation* to the context of the provided evidence.

6.3 Negation Analysis

In fact-checking, negation is one of the most difficult challenges, and many state-of-the-art models are brittle against it. Thorne and Vlachos show that the winning fact-checking systems from the FEVER workshop are brittle against negations, experiencing a huge performance drop when given negated test sets, up to absolute -29% in accuracy. Therefore, we also conduct analysis regarding the negation handling of our proposed methods by augmenting our dataset with negated examples.

Template-based Data Negation We create our negated dataset by replacing all the auxiliary verbs (e.g., is, can) with their corresponding negated forms (e.g., is not, can not), and vice versa. We apply this approach to the Covid-Scientific dataset and obtain a new version that contains {original-



(a) Covid-Scientific (b) Covid-Social (c) FEVER

Figure 2: Precision at top-k. Blue-colored marks indicate precision at each of k when claims are assigned with perplexity scores from GPT2-base to rank claims in reverse order (i.e., higher to lower score). Orange-colored marks indicate mean precision value of 10 trials when random scores are assigned to each claim to rank.

Shot #	Test Set	Models	Acc	F1-Macro
2	Original	$\texttt{RoBERTa}_{\texttt{ft}}$	52.66%	34.29%
		$\texttt{PPL}_{\texttt{GPT2-B}}$	**66.75%**	**64.39%**
	Negation-Augmented	$\texttt{RoBERTa}_{\texttt{ft}}$	46.75%	31.66%
		$\texttt{PPL}_{\texttt{GPT2-B}}$	**52.98%**	**50.99%**

Table 6: Negation Analysis - Comparison between fine-tuned RoBERTa baseline classifier ($\texttt{RoBERTa}_{\texttt{ft}}$) and our perplexity-based classifier ($\texttt{PPL}_{\texttt{GPT2-B}}$) on original Covid-Scientific dataset and its negation-augmented version in 2-shot setting.

sentence (S_{original}), negated-sentence (S_{negated})} pairs. Note that the evidence is kept the same, but the veracity label of S_{original} is negated (i.e., `Supported` is negated to `Unsupported` and vice versa). To illustrate with an example, S_{original} ={"claim": "5g helps covid-19 spread.", "evidence": $evidence_1$, "label": `Unsupported`} is negated into S_{negated} ={"claim": "5g does **not** help covid-19 spread.", "evidence": $evidence_1$, "label": `Supported`}.

Q1: Can the LM distinguish negation? We use the new augmented Covid-Scientific dataset to investigate whether the LM manages to differentiate between the original-sentence S_{original} and negated-sentence S_{negated}. The average of the absolute difference between the perplexities assigned to S_{original} and S_{negated} is 122 and the maximum absolute difference value is 2800.

Q2: Performance on negation-augmented dataset? We evaluate the performance of the perplexity-based classifier ($\texttt{PPL}_{\texttt{GPT2-B}}$) on the "negation-augmented" Covid-Scientific dataset in reference to its original. Unsurprisingly, $\texttt{PPL}_{\texttt{GPT2-B}}$ does experience a drop in performance of 13.77% and 13.40% in accuracy and F1-Macro, respectively. However, it still outperforms the fine-tuned $\texttt{RoBERTa}_{\texttt{ft}}$ baseline, the best performing baseline in the 2-shot setting, as shown in Table 6.

6.4 Comparison with existing FEVER System in Few-shot Setting

For all three tasks, we compare our perplexity models against different fine-tune baselines in Section 5.4. Unlike two newly proposed COVID-19-related tasks, FEVER is a well-established task studied by many existing works. In order to understand how our perplexity-based method compares against the literature, we conduct an additional experiment with the publicly available system from the runner-up team of the FEVER workshop, HexaF (Yoneda et al., 2018b).

We fine-tune HexaF's veracity classification modules in few-shot settings. In the 2-shot settting, HexaF shows accuracy of 49.99% and F1-Macro score of 33.33%. In the 50-shot settting, it shows accuracy of 53.53% and F1-Macro score of 49.27%. In general, machine learning models require sufficient amounts of training data, and this "sufficient amount" normally differs depending on the model being used. However, as demonstrated earlier in our main experimental results (Section 5.4), $2 \sim 50$ samples are insufficient data to properly train one of the winning fact-checking systems.

6.5 Potential Application: Ranking of Candidate Claims for Fact-Checking

Here, we discuss another way of leveraging the evidence-conditioned perplexity score. It can be

used for prioritizing false-claim candidates for human fact-checkers, instead of doing hard prediction on the veracity of the given claims. By ranking the claims-to-be-fact-checked in descending order of perplexity, we can increase the chance that the first k claims checked by a human fact-checker are `Unsupported` false claims. This will be beneficial since fact-checkers can efficiently allocate their time and resources on fact-checking claims that are more likely to be false and harmful to society.

In Figure 2, we compare the precision at the top-k (P@k) between the perplexity-based ranking and random-score-based ranking. We can view P@k to measure how many `Unsupported` pieces are prioritized in the first k of the ranked claims. Across all datasets, perplexity-based ranking (blue marks) exhibits higher precision scores over random-score-based ranking (orange marks). Moreover, for both Covid-Scientific and Covid-Social, our P@k is over 80% for all k values.

7 Future Research Directions

In this work, we conduct the FEVER experiments in a binary set-up to keep all the experimental settings consistent across all three datasets. However, the original FEVER task has three classes – Support, Refute, and Not Enough Info (NEI). Since the distinction between NEI and Refute cases is also an important problem, it would be important future work to extend our binary-class setting to the three-class setting.

Moreover, we believe our method can easily be augmented into other existing approaches, for instance, leveraging the perplexity score in the final step of the FEVER fact-checkers as additional input. It would be a useful future direction to explore and discover the most effective way of incorporating the perplexity-based approach into other existing fact-checking systems.

8 Conclusion

In this paper, we propose a novel way of leveraging the perplexity score from LMs for the few-shot fact-checking task. Through experimental analysis from an ablation study to the discussion of potential applications, we further explore and evaluate the capability of the perplexity score to act as an indicator of unsupported claims. We hope our proposed approach encourages future research to continue developing LM-based methodologies as well as the few-shot approach for fact-checking. By doing so, our community can move towards a data-efficient approach that is not constrained by the requirement of a large labeled dataset.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 85–90, Brussels, Belgium. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

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.

Maram Hasanain, Reem Suwaileh, Tamer Elsayed, Alberto Barrón-Cedeno, and Preslav Nakov. 2019. Overview of the clef-2019 checkthat! lab: Automatic identification and verification of claims. task 2: Evidence and factuality. In *CLEF (Working Notes)*.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. 2020. DeSePtion: Dual sequence prediction and adversarial examples for improved fact-checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 8593–8606, Online. Association for Computational Linguistics.

Georgi Karadzhov, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, and Ivan Koychev. 2017. Fully automated fact checking using external sources. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 344–353, Varna, Bulgaria. INCOMA Ltd.

Nayeon Lee, Belinda Z. Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? In *Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER)*, pages 36–41, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2561–2571, Florence, Italy. Association for Computational Linguistics.

Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Language models as few-shot learner for task-oriented dialogue systems.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6859–6866.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. Bert is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised qa. *ArXiv*, abs/1911.03681.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the truth lies: Explaining the credibility of emerging claims on the web and social media. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 1003–1012, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018a. Credeye: A credibility lens for analyzing and explaining misinformation. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 155–158, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018b. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham J Barezi, and Pascale Fung. 2020. Caire-covid: A question answering and multi-document summarization system for covid-19 research. *arXiv preprint arXiv:2005.03975*.

James Thorne and Andreas Vlachos. 2019. Adversarial attacks against fact extraction and verification. *arXiv preprint arXiv:1903.05543*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Santosh Tokala, Vishal G, Avirup Saha, and Niloy Ganguly. 2019. AttentiveChecker: A bi-directional attention flow mechanism for fact verification. 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 2218–2222, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008.

Lianwei Wu, Yuan Rao, Xiong Yang, Wanzhen Wang, and Ambreen Nazir. 2020. Evidence-aware hierarchical interactive attention networks for explainable claim verification. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 1388–1394. International Joint Conferences on Artificial Intelligence Organization.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32, pages 5753–5763. Curran Associates, Inc.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018a. UCL machine reading group: Four factor framework for fact finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.

Takuma Yoneda, Jeff Mitchell, Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018b. UCL machine reading group: Four factor framework for fact finding (HexaF). In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 97–102, Brussels, Belgium. Association for Computational Linguistics.

Wanjun Zhong, Jingjing Xu, Duyu Tang, Zenan Xu, Nan Duan, Ming Zhou, Jiahai Wang, and Jian Yin. 2020. Reasoning over semantic-level graph for fact checking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6170–6180, Online. Association for Computational Linguistics.

Jie Zhou, Xu Han, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2019. GEAR: Graph-based evidence aggregating and reasoning for fact verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 892–901, Florence, Italy. Association for Computational Linguistics.