A Vector-Based Approach to Few-Shot Veracity Classification for Automated Fact-Checking

Anonymous ACL submission

Abstract

As progress on automated fact-checking continues to be called, veracity classification has gained more attention. It is the task of predicting the veracity of a given claim by comparing it with retrieved pieces of evidence. One of the challenges for this task is to obtain manual annotations for large datasets, especially when it comes to new domains for which labelled data is unavailable in the first instance. In this paper, we describe a vector-based approach that achieves significant performance improvement on veracity classification in few-shot settings. Performance is compared with two competitive baselines: (1) fine-tuning BERT / RoBERTa, and (2) the state-of-the-art few-shot veracity classification approach leveraging language model perplexity with thresholds. Our approach first utilises sentence-BERT to get sentence vectors of claims and evidences. We then create a relation vector for each claim-evidences pairs, by applying absolute operation on their vector offsets. Experiments show significant improvements over the baselines.

1 Introduction

Automated fact-checking is attracting an increasing amount of attention. Despite the advances done in the task by proposing and making use of state-of-the-art natural language processing (NLP) models, the dominant approach generally requires large amount of data and/or involves training big language models. However, methods that have low demand on labelled data and are capable of being implemented and deployed fast are particularly desired in practice. Collecting a large dataset is expensive, time-consuming and may be unrealistic when time and resources are limited. For example, the current COVID-19 pandemic has triggered a remarkable amount of online misinformation (Saakyan et al., 2021). To combat the rapid spread of ongoing misinformation on new and emerging topics, fact-checkers cannot wait for the large scale datasets to become available and then train a large model in a post hoc manner. Furthermore, misinformation constantly shifts topics, resulting in collected datasets and trained models going outdated fast.

Therefore, an effective approach that can perform well by only using a small amount of data is particularly helpful in real-world settings intending to combat misinformation. We focus on the task of veracity classification, which is the task of assessing claim veracity with retrieved evidences (Thorne et al., 2018; Wadden et al., 2020; Lee et al., 2021). It is dominantly tackled as a label prediction task: given a claim c and a set of evidences e, predict the veracity label for the claim c out of "*Support*", "*Refute*" and "*NoInfo*".

This paper presents a novel and effective approach on doing veracity classification with very limited data, i.e. as little as approximately 10 samples per veracity class. We first utilise sentence BERT (Reimers and Gurevych, 2019) to get sentence vectors for the claim and its evidences and then calculate the absolute offset of the two vectors as a relation vector, which may ultimately indicate one of $[\![support]\!]$, $[\![refute]\!]$ or $[\![neutral]\!]$. Each relation vector corresponds to a veracity label. Figure 1 provides an illustration. We obtain improved relation vectors by averaging the vectors of all fit samples, which is comparably accessible and efficient and involves no gradient update. During inference, we calculate the relation vector for the claim and evidences at hand and compare its euclidean distance among the candidate relation vectors to determine the veracity.

We compare performance with two baseline systems: (1) finetuning BERT (Devlin et al., 2019)/RoBERTa (Liu et al., 2019) models to do label prediction; (2) using perplexity from BERT/GPT (Radford et al., 2019) models with a threshold to do binary classification (Lee et al., 2021). The latter is selected for being the current

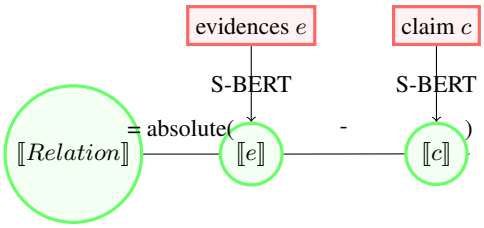

Figure 1: Illustration of our vector-based approach.

state-of-the-art model on few-shot veracity classification. Experiments show that our approach achieves significant improvements over the baselines in few-shot settings.

Our main contributions include the following:

- We achieve overall better performance on few-shot veracity classification than both baseline approaches.
- Our approach features simplicity and effectiveness, which brings low demand on labelled data and computing resources. In particular, it involves no model training.
- Our approach is flexible and can be easily adapted to any pairwise classification tasks. It also doesn't have explicit limits on the number of classes, while the previous SOTA approach (Lee et al., 2021) is restricted to binary classification.

2 Related Work

The literature on automated fact-checking has witnessed a surge over recent years, with various datasets published (Thorne et al., 2018; Chen et al., 2019; Augenstein et al., 2019; Ostrowski et al., 2020; Kotonya and Toni, 2020; Sathe et al., 2020; Wadden et al., 2020; Schuster et al., 2021; Diggelmann et al., 2021; Saakyan et al., 2021; Aly et al., 2021) and novel systems proposed (Mithun et al., 2021; Samarinas et al., 2021; Bekoulis et al., 2021).

However, following the current general trend in NLP, researchers have primarily focused on collecting and utilising large-scale datasets with lengthy pipelines whose cores are large language models which are computationally expensive, requiring resources that are not accessible to everyone. When dealing with veracity classification, most recent systems fine-tune a large pre-trained language model to do three-way label prediction, including VERISCI (Wadden et al., 2020), VERT5ERINI (Pradeep et al., 2020), ParagraphJoint (Li et al., 2021).

To the best of our knowledge, few-shot veracity classification is not well-studied. Previous efforts only experimented with perplexity-based binary veracity classification. Lee et al. (2021) hypothesised that evidence-conditioned perplexity score from language models are helpful for assessing claim veracity. They explored using perplexity scores with a threshold th to determine claim veracity into *"supported"* and *"unsupported"*: if the score is lower than the threshold th, it is classified as *"unsupported"* and otherwise *"supported"*. This perplexity-based approach has achieved better performance on few-shot binary classification than fine-tuning a BERT model. However, it is facing severe challenges when dealing with multiple classes; for example, it is not capable of tackling the veracity classification task in three-way settings involving *"Support"*, *"Refure"* and *"NoInfo"*.

3 Methodology

Our method roots in formal semantics and is inspired by one of the most well-known equations in NLP. Using a word2vec (Mikolov et al., 2013) word embedding models, researchers have achieved embedding representations that capture sufficient context to satisfy the following equation:

$$[\![king]\!] - [\![man]\!] + [\![woman]\!] = [\![queen]\!] \quad (1)$$

Despite how elegant the above equation is, direct applications on solving NLP tasks seem challenging. However, if we create a $[\![DIFF]\!]$ vector to store the elegantly captured semantic differences, we may transform it into the following:

$$[\![DIFF]\!] = [\![king]\!] - [\![man]\!] = [\![queen]\!] - [\![woman]\!] \quad (2)$$

With recent advances on efficiently applying BERT models to get sentence-level representations (Reimers and Gurevych, 2019), we may easily scale it up to capture semantic differences at the sentence level. For a sentence pair x that has $sentence_{x_a}$ and $sentence_{x_b}$, we have :

$$[\![DIFF_x]\!] = [\![sentence_{x_a}]\!] - [\![sentence_{x_b}]\!] \quad (3)$$

In the context of veracity classification that compares a claim with evidences, we are given multiple labelled sentence pairs that belong to the same class, and we can expect to obtain similar $[\![DIFF]\!]$ vectors for different instances within a class. We may then obtain the average of $[\![DIFF]\!]$ vectors

within a class, i.e., a $[\![\overline{DIFF}]\!]$ vector for each class following equation 4.

$$[\![\overline{DIFF}]\!] = \frac{1}{n}\sum_{i=1}^{n}([\![sentence_{i_a}]\!] - [\![sentence_{i_b}]\!]) \quad (4)$$

As shown above, a vector that stores pairwise semantic difference is straightforward to obtain. Furthermore, these $[\![DIFF]\!]$ vectors have great application potentials on doing pairwise classifications, especially in low-resource settings. As long as the number of the classes is manageable, we propose that we can utilise the average vectors above to do efficient classification with very few samples and very limited computing resources.

This paper demonstrates the application on the task of veracity classification.

Note that the calculated $[\![DIFF]\!]$ vectors may contain much more information than our target classification labels depending on the task. For instance, the semantic differences between the evidence "*Soul Food is a 1997 American comedy-drama film produced by Kenneth " Babyface " Edmonds , Tracey Edmonds and Robert Teitel and released by Fox 2000 Pictures.*" and the claim "*Fox 2000 Pictures released the film Soul Food.*" is the semantic meaning of the target relation $[\![support]\!]$ as well as additional information that is equivalent to sentence "*It is a 1997 American comedy-drama film produced by Kenneth " Babyface " Edmonds , Tracey Edmonds and Robert Teitel.*"

Therefore, we propose to further apply the absolute value function on every value of a $[\![DIFF]\!]$ vector to obtain a $[\![Relation]\!]$ vector. This is empirically tested to be effective at controlling the impact of random redundant information. In other words, we adapt Equation 4 to Equation 5 for the task of veracity classification:

$$[\![\overline{Relation}]\!] = \frac{1}{n}\sum_{i=1}^{n}(|[\![evidences_{i_a}]\!] - [\![claim_{i_b}]\!]|) \quad (5)$$

During inference, we calculate the euclidean distance between the current $[\![Relation]\!]$ vector and every target $[\![Relation]\!]$ vector, i.e., $[\![support]\!]$, $[\![refute]\!]$ and $[\![neutral]\!]$ for the task of veracity classification, and make predictions on the one that has smallest euclidean distance value.

4 Experiments

4.1 Datasets

We conduct experiments on the Fact Extraction and Verification (FEVER) (Thorne et al., 2018) and SCIFACT (Wadden et al., 2020) datasets. FEVER is one of the most well-studied large-scale datasets for automated fact-checking. It contains claims that are manually modified from Wikipedia sentences and their corresponding Wikipedia evidences. Claims are annotated into three categories: *"Support"*, *"Refute"* and *"Not Enough Info"*. SCIFACT is a smaller dataset that focuses on scientific claim verification. The claims are annotated by experts and evidences are retrieved from research paper abstracts. Similarly, claims are classified into *"Support"*, *"Contradict"* and *"Not_ Enough_Info"*.

4.2 Model implementation

We implemented our vector-based approach by utilising sentence BERT (Reimers and Gurevych, 2019) with huggingface transformers model hub (Wolf et al., 2020). Specifically, we use three variants of BERT (Devlin et al., 2019) as the base model: BERT-base, BERT-large and BERT-base-nli. We use VEC_{BERT_B}, VEC_{BERT_L} and $VEC_{BERT_{B-NLI}}$ to denote them thereafter. The first two use vanilla BERT models available from huggingface model hub with model id *bert-base-uncased* and *bert-large-uncased* respectively. The last one is a sentence BERT model that has been fine-tuned on natural language inference (NLI) tasks and is available on sentence BERT repository with model id *bert-base-nli-mean-tokens*. We include experiments with $VEC_{BERT_{B-NLI}}$ in both binary and three-way veracity classifications as natural language inference is high relevant.

4.3 Baselines

Our baseline fine-tuning approach to binary classification involves fine-tuning BERT-base, BERT-large, RoBERTa-base and XLNET-base (Yang et al., 2019), which are denoted as FT_{BERT_B}, FT_{BERT_L}, $FT_{RoBERTa_B}$ and FT_{XLNET_B} thereafter. Furthermore, the perplexity-based approach involves GPT-2 and BERT models, which are denoted as PPL_{BERT_B} and PPL_{GPT_B}. Please see implementation details in Lee et al. (2021).

We finetune the following models as baselines for our three-way veracity classification: BERT-base, BERT-large, RoBERTa-base and RoBERTa-large. We utilise huggingface transformers library

(Wolf et al., 2020) for easily finetuning these models, with model id being *bert-base-uncased*, *bert-large-uncased*, *roberta-base* and *roberta-large* respectively.

4.4 Results

Experiments are conducted in both binary and three-way settings for FEVER and three-way for SCIFACT. Binary classification on FEVER enables direct comparison with the SOTA model, i.e. perplexity-base approach, while the three-way classification on both FEVER and SCIFACT dataset demonstrate the further potential of our approach on both datasets.

4.4.1 Results on FEVER binary classification

We first report performance on the FEVER dataset in binary setting for direct comparison with the SOTA model. To do this, and in line with previous work, we merge *"Refute"* and *"Not Enough Info"* into *"Unsupport"*. This is done for comparing performance with the perplexity-based approach as it is only available in this particular setting.

Following the baseline paper (Lee et al., 2021), we only use the test set of FEVER dataset, as it already has lots of claims. Specifically, we randomly sample 3333 instances out of *"Support"* class, 1666 instances and 1667 instances out of the *"Not Enough Info"* class and the *"Refute"* class respectively. Samples of the latter two classes are treated as *"Unsupport"*. We sample n shots, i.e. n instances per class, as the training set for the fine-tuning approach and as the fit set for the perplexity-based approach and the vector-based approach, and use the rest as test set to evaluate performance.

Table 1 reports results on binary setting of FEVER dataset. Model names that start with "FT" refer to methods directly fine-tuning the pre-trained language models indicated in the subscript. Model names that start with "PPL" refer to perplexity-based, state-of-the-art methods. Model names that start with "VEC" refer to our vector based approach.

As shown in the table, all of the fine-tuned models achieve very low performance in all 2-shot, 10-shot and 50-shot settings, with their accuracy peaking at around 50%, which is equivalent to a random classifier in the binary setting. PPL_{GPT_B} achieves best performance with 2 shots, but struggles to gain significant improvements with more shots. Interestingly, $VEC_{BERT_{B-NLI}}$ achieves a substantial performance boost of almost 20 points when increasing the fitset from 2 shots to 10 shots, and achieves the best overall performance in the 10-shot and 50-shot settings. Furthermore, with the same base language model $BERT_B$, vector-based approach VEC_{BERT_B} always outperforms the perplexity-based approach PPL_{BERT_B} in terms of accuracy: its accuracy is 1.15%, 14.18% and 14.29% higher in 2-shot, 10-shot and 50-shot settings respectively. Regarding macro F1, VEC_{BERT_B} is only 0.15% lower in the 2-shot setting than PPL_{BERT_B}, but 13.8% and 14.65% higher in 10-shot and 50-shot settings. Overall, the vector-based approach achieves significant improvements in 10-shot and 50-shot settings, compared to both the fine-tuning approach and the perplexity-based approach.

4.4.2 Results on FEVER three-way classification

We then report our experiments on the three-way setting with the FEVER dataset. Due to the innate limitations of the perplexity-based approach, i.e., only suitable for binary classification, we do not include it in three-way experiments. For the vector-based approach, we experiment with three base models that are available for sentence BERT to use, namely *bert-base-uncased*, *bert-large-uncased* and *bert-base-nli-mean-tokens*. The first two are vanilla BERT model while the last one is a BERT-base model that is finetuned on natural language inference tasks (Reimers and Gurevych, 2019). For the fine-tuning approach, we experiment with *bert-base-uncased*, *bert-large-uncased*, *roberta-base* and *roberta-large*.

We randomly sample 3333 instances for each class out of *"Support"*, *"Refute"* and *"Not Enough Info"*. Experiments for both the fine-tuning approach and the vector-based approach are conducted in 2-shot, 4-shot, 6-shot, 8-shot, 10-shot, 20-shot, 30-shot, 40-shot, 50-shot and 100-shot settings with 10 different random seeds: 123, 124, 125, 126, 127, 128, 129, 130, 131, 132. Due to the variability in performance of the fine-tuning approach introduced by its non-deterministic nature, we do 5 runs for each setting and report the average results. Following the baseline paper (Lee et al., 2021), We use $5e^{-6}$ for FT_{BERT_B} and $FT_{RoBERTa_B}$ as learning rate and $2e^{-5}$ for FT_{BERT_L} and $FT_{RoBERTa_L}$. All models share the same batch size of 32 and are trained for 10 epochs.

Figure 2 reports the accuracy scores for both

Model	Accuracy			Macro-F1		
# of Shots	2	10	50	2	10	50
FT_{BERT_B}	51.56	51.56	52.18	37.34	37.34	38.82
FT_{BERT_L}	50.80	50.80	51.14	36.49	36.49	39.99
$FT_{RoBERTa_B}$	50.00	50.00	50.44	33.33	33.33	38.15
FT_{XLNET_B}	49.41	49.41	49.18	44.65	44.65	48.42
PPL_{BERT_B}	52.54	57.59	57.44	41.33	57.11	56.94
PPL_{GPT_B}	**61.92**	62.82	67.48	**57.50**	57.04	64.7
VEC_{BERT_B}	53.69	71.77	71.73	41.18	70.91	71.59
VEC_{BERT_L}	52.31	69.83	70.51	39.34	69.26	70.34
$VEC_{BERT_{B-NLI}}$	55.31	**81.12**	**80.89**	44.41	**81.1**	**80.89**

Table 1: Comparison of Few-Shot Performance on binary setting of the FEVER Dataset. The baseline results are taken from the baseline paper (Lee et al., 2021).

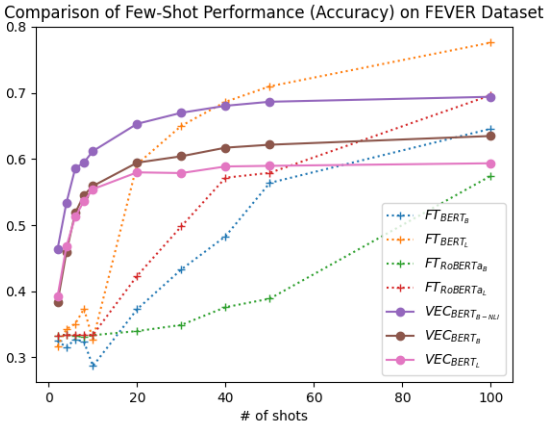

Figure 2: Comparison of Few-Shot Accuracy Performance on the FEVER Dataset.

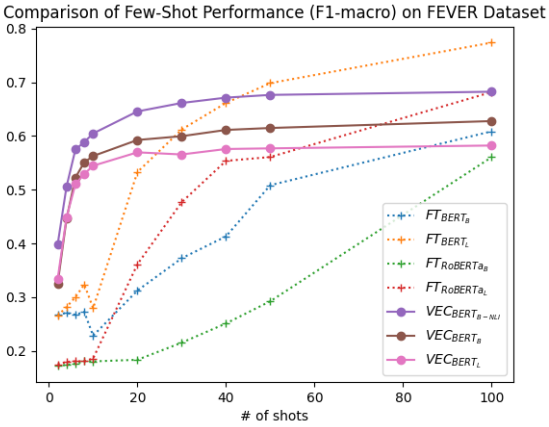

Figure 3: Comparison of Few-Shot Macro F1 Performance on the FEVER Dataset.

approaches on the three-way veracity classification task of FEVER dataset in various few-shot settings, while Figure 3 reports the macro F1 score. FT_{BERT_B}, FT_{BERT_L}, $FT_{RoBERTa_B}$ and $FT_{RoBERTa_L}$ follow our baseline approach and finetune a BERT-base, BERT-large, RoBERTa-base and RoBERTa-large model respectively. VEC_{BERT_B}, VEC_{BERT_L} and $VEC_{BERT_{B-NLI}}$ follow our proposed vector-based approach with model base choice of *bert-base-uncased*, *bert-large-uncased* and *bert-base-nli-mean-tokens* model respectively.

Both figures show consistent performance increase trend with increasing amount of data for both approaches. However, it is clear that when given under 20 shots, the vector-based approach has significant performance advantages. The vector-based approach starts its performance boost right at the beginning, but the baseline approach has a delay and only slowly start its performance boost after 10-shots. This proves the vector-based approach as effective compared to other methods, particularly when labelled data is scarce.

4.4.3 Results on SCIFACT three-way classification

Furthermore, experiments on three-way veracity classification with the SCIFACT dataset show similar trend. The SCIFACT dataset is much smaller than the FEVER dataset, with only 809 claims in the training set and 300 claims in the development set (the test set is not yet available at the time of writing as it was withheld for a shared task). We use both the train set and dev set. We randomly

sample n instances for each class out of *"Support"*, *"Contradict"* and *"Not Enough Info"* and use them as the fit/train set, with n being one of 2, 4, 6, 8, 10, 20, 30, 40, 50, 100. We randomly sample 70 instances for each class in dev set and use them to evaluate performance, as the dev set is very unbalanced with 138, 114 and 71 pairs for each class. Following the same methodology as with the FEVER dataset, we also sample with 10 different random seeds: 123, 124, 125, 126, 127, 128, 129, 130, 131, 132 and do 5 runs for each setting for the fine-tuning approach. The reported performance scores are the average results over multiple runs and multiple random samplings of seeds. We use the same hyperparameters as above.

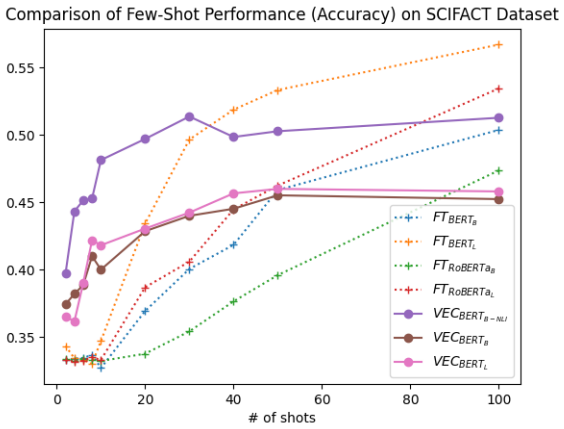

Figure 4: Comparison of Few-Shot Accuracy Performance on the SCIFACT Dataset.

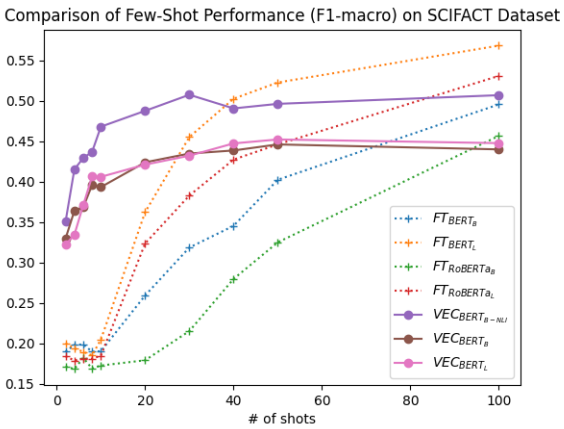

Figure 5: Comparison of Few-Shot Macro F1 Performance on the SCIFACT Dataset.

Figure 4 reports the accuracy scores for both approaches on three-way veracity classification task of SCIFACT dataset in various few-shot settings, while Figure 5 reports the macro F1 score. FT_{BERT_B}, FT_{BERT_L}, $FT_{RoBERTa_B}$ and $FT_{RoBERTa_L}$ follow our baseline approach and finetune a BERT-base, BERT-large, RoBERTa-base and RoBERTa-large model respectively. VEC_{BERT_B}, VEC_{BERT_L} and $VEC_{BERT_{B-NLI}}$ follow our proposed vector-based approach with model base choice of BERT-base, BERT-large and BERT-base-nli-mean-tokens model respectively.

Similar to the results on the FEVER dataset, consistent performance increase trends are shown with increasing amount of data for both approaches. Noticeably the absolute performance scores on the SCIFACT dataset is overwhelmingly lower than the FEVER dataset. Though with more fluctuations, the vector-based approach still starts its performance boost right at the beginning. The baseline approach has a more severe delay and experience almost no performance gain when given fewer than 10-shots. This indicates that the Vector-based approach maintains significant performance advantages in few-shot settings even with a much more difficult dataset.

4.4.4 Summary of results

These experiments have demonstrated the effectiveness of our proposed approach when doing veracity classification in few-shot settings. By making use of relation vectors, our approach is capable of doing multi-class classification with only very few samples. With only 10 shots, our approach achieves approximately 80% accuracy on binary veracity classification, which is about 30% higher than the fine-tuning approach and 20% higher than the perplexity-based approach. Furthermore, we achieve 60% accuracy on three-way veracity classification within the domain of general Wikipedia texts, while the fine-tuning approach would only achieve around 33% accuracy, which is similar to a random guess. Given the difficulty of performing veracity classification on scientific texts in the SCIFACT dataset, our approach still achieves accuracy above 40% , while the performance of the fine-tuning approach remains similar to a random guess.

5 Discussion and Future Work

5.1 Discussion

Thanks to its simplicity, our vector-based approach yields very good performance in few-shot veracity classification. By outperforming competitive base-

lines, including a state-of-the-art model and Transformers, we show the potential of our approach in scenarios where the scarcity of labelled data and/or computing resources require the use of a light-weight approach. This in turn validates our proposed methodology that averaging pairwise offsets between claims and evidences for each class can lead to meaningful vectors that can help characterise each of the classes for the veracity classification task: *"Support"*, *"Refute"* and *"NoInfo"*.

Further work on applying it to other pairwise classification tasks are likely to yield similarly exciting results. One may question the simplicity of the proposed approach. However, it is thanks to this simplicity that our method overcome the heavy reliance of advanced machine learning techniques on large amount of training data which in turn makes them struggle to work in few-shot settings.

We also aim to bring researchers' attention to the cost effectiveness of data usage. In the context of automated fact-checking, naturally occurring labelled data would only emerge slowly in a post hoc manner, due to the complexity of the task. Moreover, collected data may become outdated quickly as new events continue to occur. Our approach has demonstrated the potential for leveraging a small amount of data when being used effectively. We hope it will encourage future research to study the cost effectiveness of collecting large amount of data for simple tasks.

5.2 Future Work

Our work opens up a few directions for future research, which we discuss next.

5.2.1 Sentence Vectors

To further customise the approach to a specific task, we believe that getting better sentence vectors is a promising direction. For example, we may use BioBERT (Lee et al., 2020) as the base model other than vanilla BERT, when dealing with biomedical texts. Alternatively, a BERT model that is already finetuned on relevant tasks is likely to yield a stronger presence of the target information.

5.2.2 Relation Vectors

We proposed to apply absolute operation to get our relation vectors as it is effective to reduce the amount of preserved information. However, it has also cancelled the directionality of relation vectors. For example, the evidence "*England has the best football team in Europe.*" supports the claim "*England has the best football team in the UK.*" but the support relation doesn't stay if the claim and the evidence are swapped. Future work on preserving the directionality without losing the simplicity and effectiveness is highly desired.

5.2.3 Cross-dataset Applications

As demonstrated above, a specific relation vector for a specific domain may be easily obtained and has good performance when doing inference within the same dataset. An interesting direction is to make use of these vectors in a cross-dataset manner. For example, to obtain a $[\![support]\!]$ vector from a natural language inference dataset and use it on the task of veracity classification. The major challenge here lies in domain adaptation.

6 Conclusions

We have presented a simple but effective vector-based approach which achieves significant improvements over the baseline systems in few-shot veracity classification. It has very low demand on data quantity and computing resources. We have also demonstrated that future research on cost effectiveness of data usage is highly valued.

Acknowledgements

Omitted for blind review.

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. *arXiv:2106.05707 [cs]*. ArXiv: 2106.05707.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. *arXiv:1909.03242 [cs, stat]*. ArXiv: 1909.03242.

Giannis Bekoulis, Christina Papagiannopoulou, and Nikos Deligiannis. 2021. Understanding the Impact of Evidence-Aware Sentence Selection for Fact Checking. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 23–28, Online. Association for Computational Linguistics.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing Things

from a Different Angle:Discovering Diverse Perspectives about Claims. 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 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims. *arXiv:2012.00614 [cs]*. ArXiv: 2012.00614.

Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking: A Survey. *arXiv:2011.03870 [cs]*. ArXiv: 2011.03870.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. Publisher: Oxford Academic.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2021. Towards Few-shot Fact-Checking via Perplexity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1971–1981, Online. Association for Computational Linguistics.

Xiangci Li, Gully Burns, and Nanyun Peng. 2021. A Paragraph-level Multi-task Learning Model for Scientific Fact-Verification. *arXiv:2012.14500 [cs]*. ArXiv: 2012.14500.

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:1907.11692 [cs]*. ArXiv: 1907.11692.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

Mitch Paul Mithun, Sandeep Suntwal, and Mihai Surdeanu. 2021. Data and Model Distillation as a Solution for Domain-transferable Fact Verification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4546–4552, Online. Association for Computational Linguistics.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2020. Multi-Hop Fact Checking of Political Claims. *arXiv:2009.06401 [cs]*. ArXiv: 2009.06401.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2020. Scientific Claim Verification with VERT5ERINI. *arXiv:2010.11930 [cs]*. ArXiv: 2010.11930.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv:1908.10084 [cs]*. ArXiv: 1908.10084.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-Fact: Fact Extraction and Verification of Real-World Claims on COVID-19 Pandemic. *arXiv:2106.03794 [cs]*. ArXiv: 2106.03794.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2021. Improving Evidence Retrieval for Automated Explainable Fact-Checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91, Online. Association for Computational Linguistics.

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated Fact-Checking of Claims from Wikipedia. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6874–6882, Marseille, France. European Language Resources Association.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. *arXiv:2103.08541 [cs]*. ArXiv: 2103.08541.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. *arXiv:1803.05355 [cs]*. ArXiv: 1803.05355.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. *arXiv:2004.14974 [cs]*. ArXiv: 2004.14974.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020.

HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]*. ArXiv: 1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. 32.