Neural Re-rankers for Evidence Retrieval in the FEVEROUS Task

Mohammed Saeed, Giulio Alfarano, Khai Nguyen, Duc Pham, Raphaël Troncy and Paolo Papotti
EURECOM, France
{FirstName.LastName}@eurecom.fr

Abstract

Computational fact-checking has gained a lot of traction in the machine learning and natural language processing communities. A plethora of solutions have been developed, but methods which leverage both structured and unstructured information to detect misinformation are of particular relevance. In this paper, we tackle the FEVEROUS (Fact Extraction and VERification Over Unstructured and Structured information) challenge which consists of an open source baseline system together with a benchmark dataset containing 87,026 verified claims. We extend this baseline model by improving the evidence retrieval module yielding the best evidence F1 score among the competitors in the challenge leaderboard while obtaining an overall FEVEROUS score of 0.20 (5^{th} best ranked system).

1 Introduction

The volume of potentially misleading and false claims has surged with the increasing usage of the web and social media. No barriers exist for publishing information which make anyone capable of diffusing false or biased claims while reaching large audiences with ease (Baptista and Gradim, 2020). One approach of dealing with this ordeal is computational fact-checking (Wu et al., 2014), where the automation of the verification pipeline or parts of it is flourishing due to advances in natural language processing (Nakov et al., 2021; Saeed and Papotti, 2021). Along these lines, several datasets and fact-evaluation algorithms have been proposed (Kotonya and Toni, 2020).

In this paper, we report on our effort in tackling the FEVEROUS challenge (Aly et al., 2021). The provided dataset consists in a set of textual claims verified against evidence retrieved from a corpus of English Wikipedia pages. The claims are labeled as *supported*, *refuted* or *NEI* (Not Enough Information). Evidence can be unstructured (such as sentences) or structured (such as general tables, infoboxes, lists, etc.). The task is to return the right label with the correct evidence. The baseline model is divided in two main parts: an evidence retrieval part and a verdict prediction part. The evaluation is performed through the so-called FEVEROUS score which is computed considering both the correct retrieval of the evidence and the correct label predictions. In this paper, we propose an enhanced version of this baseline model that focuses on the retrieval component through a re-ranking process of pages, resulting in a more precise model.

In the remainder of this paper, we first describe briefly the challenge task and the supplied data, and we detail our extension (Section 2). We then provide experimental results obtained on the development dataset and we discuss observations on analyzed errors (Section 3). We conclude with a discussion of other research directions that can be applied to improve results for the FEVEROUS task (Section 4).

2 Method

We begin by reviewing the given baseline, and then propose an extension to it that improved the precision and recall of the page-retrieval module in exchange for more computation time.

2.1 FEVEROUS Baseline

In FEVEROUS (Aly et al., 2021), the aim is to find out the veracity of a claim c. This is done by: (i) acquiring a set of evidence E which could contain sentences extracted from a Wikipedia page, or cell(s) from a Wikipedia table, and (ii) predicting a label $y \in$ {Supports, Refutes, NEI}.

The proposed baseline is simple yet competitive (Aly et al., 2021). For (i), a combination of entity-matching and TF-IDF scoring are used to identify the most prominent Wikipedia pages (Chen et al., 2017). k pages are selected by matching entities extracted from the claim to Wikipedia pages.

If needed, remaining pages are identified using TF-IDF matching between the claim and the introductory sentence of the page. Given the extracted Wikipedia pages, sentences are scored through a dot product with the claim in the TF-IDF space, where the top l sentences are retrieved. Similarly, the top q tables are extracted where the TF-IDF vector of the table title is used to represent a table. The tables are then linearized, pre-processed to respect the input-size limit of the classifier (Oguz et al., 2020), and then used, alongside the claim, to fine-tune a RoBERTa model (Liu et al., 2020) on a binary token classification task.

For (ii), given the retrieved evidence, the final verdict is predicted using a RoBERTa model with a sequence-classification layer which is fed with sequentially concatenated claim and evidences as input. The model has been trained on a set of labelled claims (71,291 samples) with their associated evidence.

2.2 Proposed Extension

It is clear that to enhance the system, evidence retrieval should be a top priority as identifying the correct evidence is crucial for the verdict predictor to function properly. We focus on enhancing the identification of Wikipedia pages by utilizing advances in the information retrieval (IR) community where neural ranking models have been proposed for better data retrieval (Mitra et al., 2016; Hui et al., 2018).

A simple IR pipeline comprises a two-stage reranking process where: (a) first, a large number of documents to a given query are retrieved from a corpus using a standard mechanism such as TF-IDF or BM25; (b) second, the documents are scored and reranked using a more computationally-demanding method. Given that neural ranking methods have shown success in the IR community (Guo et al., 2019), we used this method as part of our extension.

For (a), we use the current page-retriever based on entity-matching and TF-IDF to retrieve a higher number of pages. For (b), the re-ranker model provides a score s_i indicating how relevant a page p_i is to an input claim c. The re-ranker is based on a pre-trained BERT model (Devlin et al., 2019) that is fine-tuned on the passage re-ranking task of the MS MACRO dataset (Nguyen et al., 2016) to minimize the binary cross-entropy loss:

$$L = -\sum_{i \in I^+} log(s_i) - \sum_{i \in I^-} log(1 - s_i) \qquad (1)$$

where I^+ and I^- are the set of indices of the relevant and non-relevant pages respectively in the top-1,000 MS MACRO documents retrieved with BM25 (Nogueira and Cho, 2019). We designate a page re-ranker model as a function $PR(m)$ which take a set of relevant pages for a claim, which are usually scored by a less-computationally demanding method such as TF-IDF to limit the set of candidates, scores them, and outputs the top m pages.

3 Experiments

Model. We rely on the cross-encoder model fine-tuned on the MS MARCO Passage Ranking task (Reimers and Gurevych, 2019) provided on the Hugging Face model hub. We feed the claim with every extracted page into the re-ranker model to obtain scores used to re-rank the respective pages.
Settings. We set $k = 150$ and $m = 5$ where 150 pages are first extracted through entity-matching+TF-IDF, scored with the re-ranker model, and then the top-5 pages for each claim are extracted. The remainder of the pipeline remains intact ($l = 5$, $q = 3$). We designate this pipeline by $BLpage(150) \rightarrow PR(5) \rightarrow tfidf(5, 3)$. Our code can be found at `https://gitlab.eurecom.fr/saeedm1/eurecom-fever`.
Results. We see improvement with the page re-ranker as the coverage of documents has increased compared to the baseline. Hence, for $k = 5$, the retriever without the re-ranker achieves a document coverage of 69% on the dev set, while the addition of the re-ranker enhances the coverage to around 79%, which, in turn, improves the FEVEROUS and F1 scores, compared to the initial baseline (Table 1, in bold). While the page re-ranker improves the document coverage, we do not observe pronounced improvements on the system as a whole. Even with a better page retriever, an increase in FEVEROUS and F1 scores requires improvements also in the sentence and cell evidence retrievers.

Although more time-demanding, the re-ranker gives more subtle results than an entity-matching approach (Aly et al., 2021). For example, given the following excerpt of a claim: “**Family Guy is an American animated sitcom that features five main voice actors [...] and has appeared in 22 (out of 349) episodes [...] that has appeared**

	FS	**LA**	**EP**	**ER**	**E-F1**
$BL(5,5,3)$	0.19	0.54	0.12	0.29	0.17
$BL(5,5,3)_{full}$	0.186	0.533	0.119	0.289	0.168
$BLpage(50) \rightarrow PR(5) \rightarrow tfidf(5,3)$	0.129	0.468	0.120	0.201	0.151
$BLpage(150) \rightarrow PR(5) \rightarrow tfidf(5,3)$	**0.218**	**0.548**	**0.145**	**0.339**	**0.203**
$BLpage(150) \rightarrow PR(5) \rightarrow BM25(5,3)$	0.205	0.550	0.127	0.321	0.182
$BLpage(150) \rightarrow PR(5) \rightarrow SR(5) \rightarrow tfidf_{table}(3)$	0.184	0.501	0.130	0.283	0.179

Table 1: Results on the dev set showing the FEVEROUS Score (FS), the Label Accuracy (LA), the Evidence Precision (EP), the Evidence Recall (ER), and the Evidence F1-score (E-F1) of the different system variants.

in 90 episodes.", one can directly see that the retrieved pages should be related to the series "**Family Guy**". The baseline fails to predict the correct page `Family Guy`, and instead matches with entities such as `Guy` and `American`, and Wikipedia pages for numbers such as `90` and `22`. Additionally, some pages retrieved with TF-IDF do not relate to the claim at hand: `John Manwood (MP)` and `John Manwood`. The page re-ranker, on the other hand, manages to get the correct page in the top-5 predictions, where all other predictions are related: `List of Family Guy guest stars`, `List of Family Guy episodes`, `Blue Harvest` (an episode from the TV series), `List of Family Guy cast members`, and `Family Guy`. Finally, we observe that the entity-matching process is brittle and fails to match the sub-string "Angela Santomero" to the page `Angela C. Santomero` as it only performs exact string matching.

There are some cases where entity-matching+TF-IDF outperformed the re-ranker: some of those are cases where the Wikipedia page content is small and does not bring much benefit on a semantic level and this is where TF-IDF works better. We observe that we tend to miss the correct page when there are several pages who share similar semantics. For example, given the claim: "**Seven notable animated television series, including Super Why!, a children's educational show created by Angela C. Santomero and Samantha Freeman Alpert, Phineas and Ferb and WordGirl, were released in September 2007.**", the page re-ranker retrieves TV shows that are produced `Angela C. Santomero`). However, the correct page `Phineas and Ferb` does not appear in the top-5 predictions, and other pages take the lead, whereas the baseline can identify the correct page by entity matching, although its predictions are not as coherent as those of the page re-ranker.

k	**FEVEROUS Score**	**Time(mins)**
200	0.216	140
300	0.224	210
500	0.219	345

Table 2: Results on the dev set showing the FEVEROUS Score and the recorded time for the re-ranking processing for varying values of k.

Other Attempts. We have experimented with varying the number of extracted pages k. We measure also the time taken for re-ranking. Table 2 shows the results. We observe that increasing the number of pages to extract does not always increase the FEVEROUS score, as more candidate pages act as distractors to the other modules in the pipeline.

We have attempted to perform other extensions to the system that we describe below (Table 1).

Firstly, we specified the re-ranking system to extract less pages (50), but it worsened the scores. This configuration is defined as $BLpage(50) \rightarrow PR(5) \rightarrow tfidf(5,3)$.

Furthermore, we applied the same re-ranking approach at the sentence level. After obtaining 150 pages from the page re-ranker, we continue to retrieve all sentences from every page and re-rank them using the same passage re-ranking model (Nguyen et al., 2016), ($BLpage(150) \rightarrow PR(5) \rightarrow SR(5) \rightarrow tfidf_{table}(3)$). However, despite great outputs of page re-ranker, we could not obtain better results from sentences re-ranker than TF-IDF.

Regarding how relevant sentences and tables are chosen as evidence, apart from TF-IDF + Cosine Similarity, we also experimented with the Okapi BM25 scoring function (Robertson et al., 1995). This is applied after the pages are re-ranked, $(BLpage(150) \rightarrow PR(5) \rightarrow BM25(5,3))$. Surprisingly, although BM25 is generally preferred for document retrieval, in our case, it did not lead to better results compared to TF-IDF. One possible cause might lie in text preprocessing, as we did not fully explore different combinations of preprocessing functions.

Lastly, we attempted to improve the verdict predictor by (i) fine-tuning the verdict classifier on the full training dataset ($BL(5,5,3)_{full}$) and by (ii) utilizing other pre-trained models that are either larger or were pre-finetuned on a NLI dataset. However, we did not observe significant improvements from them since their performance on the dev set was either on par or slightly worse than the baseline model signaling that the focus enhancing of the second part of the system requires more significant changes.

4 Conclusion and Future Directions

In this work, we have proposed the inclusion of a neural re-ranker model as a refinement step after standard methods such as TF-IDF. While being more intensive on the computational side, we do see improvements on the document-retrieval side where results are more sound. There are of course more directions that are worth exploring to improve the results further.

Sentence retrieval could be improved by incorporating a pre-trained neural network that performs semantic matching between the claims and the sentences. One instance of such models is where the text sequences are encoded, then passed through an alignment layer that computes aligned representations of the encoded input sequences, followed by a matching layer that performs the semantic matching (Nie et al., 2018). Such models have been applied on the FEVER dataset (Thorne et al., 2018) and have been shown to outperform the TF-IDF approach (Nie et al., 2018).

Cell retrieval could be enhanced by utilizing pre-trained models over tables that outperform pre-trained models over text (Herzig et al., 2021). Several systems that exploit table structure have been proposed for the task of fact-checking a claim over a table. However, not all of them can be used in every setting as each system holds different attributes and dimensions that need to be comprehended to better integrate them in certain tasks (Saeed and Papotti, 2021). For example, some systems such as SCRUTINIZER (Karagiannis et al., 2020) are dependent on the table-schema and would not benefit in the FEVEROUS scenario where tables have varying schema. Yet, other systems such as TAPAS (Herzig et al., 2020) are schema-independent and can be fine-tuned on the available FEVEROUS dataset to provide a score for a given table, thus acting as a table ranker module. Some of these systems can be even directly trained on the data, to get domain-specific models. Once the tables have been identified, a classifier can be trained on top of models that output cell representations of a table, such as TaBERT (Yin et al., 2020) and TURL (Deng et al., 2020), to extract the key cells for verdict prediction. Also, fine-tuning the re-ranker models on the given data is a viable approach. Finally, more sophisticated entity matching algorithms could have been explored to avoid the "exact match" issues that we observed with the baseline's entity matching (C. et al., 2018).

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: fact extraction and verification over unstructured and structured information. *CoRR*, abs/2106.05707.

João Pedro Baptista and Anabela Gradim. 2020. Understanding fake news consumption: A review. *Social Sciences*, 9(10).

Paul Suganthan G. C., Adel Ardalan, AnHai Doan, and Aditya Akella. 2018. Smurf: Self-service string matching using random forests. *Proc. VLDB Endow.*, 12(3):278–291.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*.

Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2020. Turl: Table understanding through representation learning. *Proc. VLDB Endow.*, 14:307–319.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, pages 4171–4186.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2019. A deep look into neural ranking models for information retrieval. *CoRR*, abs/1903.06902.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 512–519, Online. Association for Computational Linguistics.

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.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2018. Co-pacrr: A context-aware neural ir model for ad-hoc retrieval. In *Proceedings of Web Search and Data Mining 2018*.

Georgios Karagiannis, Mohammed Saeed, Paolo Papotti, and Immanuel Trummer. 2020. Scrutinizer: A mixed-initiative approach to large-scale, data-driven claim verification. *Proc. VLDB Endow.*, 13(11):2508–2521.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A robustly optimized BERT pretraining approach.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2016. Learning to match using local and distributed representations of text for web search. *CoRR*, abs/1610.08136.

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *IJCAI*, pages 4826–4832. ijcai.org.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2018. Combining fact extraction and verification with neural semantic matching networks. *CoRR*, abs/1811.07039.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Barlas Oguz, Xilun Chen, Vladimir Karpukhin, Stan Peshterliev, Dmytro Okhonko, Michael Sejr Schlichtkrull, Sonal Gupta, Yashar Mehdad, and Scott Yih. 2020. Unified open-domain question answering with structured and unstructured knowledge. *CoRR*, abs/2012.14610.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST.

Mohammed Saeed and Paolo Papotti. 2021. Fact-checking statistical claims with tables. *IEEE Data Eng. Bull.*, 44(3).

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.

You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2014. Toward computational fact-checking. *Proc. VLDB Endow.*, 7(7):589–600.

Pengcheng Yin, Graham Neubig, Wen tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Annual Conference of the Association for Computational Linguistics (ACL)*.