The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) Shared Task

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos
Department of Computer Science and Technology
University of Cambridge
{rmya2,zg283,mss84,jt719,av308}@cam.ac.uk

Christos Christodoulopoulos
Amazon Alexa
chrchrs@amazon.co.uk

Oana Cocarascu
Department of Informatics
King's College London
oana.cocarascu@kcl.ac.uk

Arpit Mittal
Facebook
arpitmittal@fb.com

Abstract

The Fact Extraction and VERification Over Unstructured and Structured information (FEVEROUS) shared task, asks participating systems to determine whether human-authored claims are SUPPORTED or REFUTED based on evidence retrieved from Wikipedia (or NOTENOUGHINFO if the claim cannot be verified). Compared to the FEVER 2018 shared task, the main challenge is the addition of structured data (tables and lists) as a source of evidence. The claims in the FEVEROUS dataset can be verified using only structured evidence, only unstructured evidence, or a mixture of both. Submissions are evaluated using the FEVEROUS score that combines label accuracy and evidence retrieval. Unlike FEVER 2018 (Thorne et al., 2018a), FEVEROUS requires partial evidence to be returned for NOTENOUGHINFO claims, and the claims are longer and thus more complex. The shared task received 13 entries, six of which were able to beat the baseline system. The winning team was "Bust a move!", achieving a FEVEROUS score of 27% (+9% compared to the baseline). In this paper we describe the shared task, present the full results and highlight commonalities and innovations among the participating systems.

1 Introduction

Automated fact verification has become an important field of research, as fact-checkers and journalists are facing an even-increasing volume of claims to verify (Thorne and Vlachos, 2018). This task has been explored by the NLP community through forums and shared tasks such as CLEF CheckThat! (Nakov et al., 2021), SemEval (Wang et al., 2021) and FEVER (Thorne et al., 2018b), as well as a number of datasets aimed at modelling parts of the task (Karadzhov et al., 2017; Wang, 2017; Augenstein et al., 2019; Chen et al., 2020; Gupta et al., 2020).

While these previous works focus on claims that are verified against a single type of evidence, such as text *or* structured information, the new FEVEROUS dataset (Aly et al., 2021) we study in this shared task requires the models to reason about *both* types of evidence. This helps better approximate real-world fact checking, where both the claims and the sources of evidence are more complex in nature.

FEVEROUS models the task of Fact Extraction and VERification Over Unstructured and Structured information, overcoming some limitations of the previous FEVER dataset (Thorne et al., 2018a), which only considers text as evidence, and improving the quality of the annotations as well as removing known biases from the dataset. FEVEROUS contains 87,026 new claims which are more complex (25.3 word/claim on average compared to 9.4 for FEVER), and a larger pool of evidence (tables, lists, and sentences from the entirety of Wikipedia), bringing us closer to real-world scenarios, while maintaining the experimental control of an artificially designed dataset.

This paper presents a short description of the task and dataset, the final test phase leaderboard, and a summary of the submissions with a comparison to previous FEVER shared tasks, an analysis of current challenges and a discussion around interesting research directions for this task. The shared task received 13 entries in total, with the winning team, *"Bust a move!"*, achieving a score of 27%, 9 percentage points higher than the baseline system we released. While considerable progress was made by the participants of the task, there are still plenty of opportunities for systems to improve. We will leave the scoring system open to allow future work to build upon the advances made in this shared task.

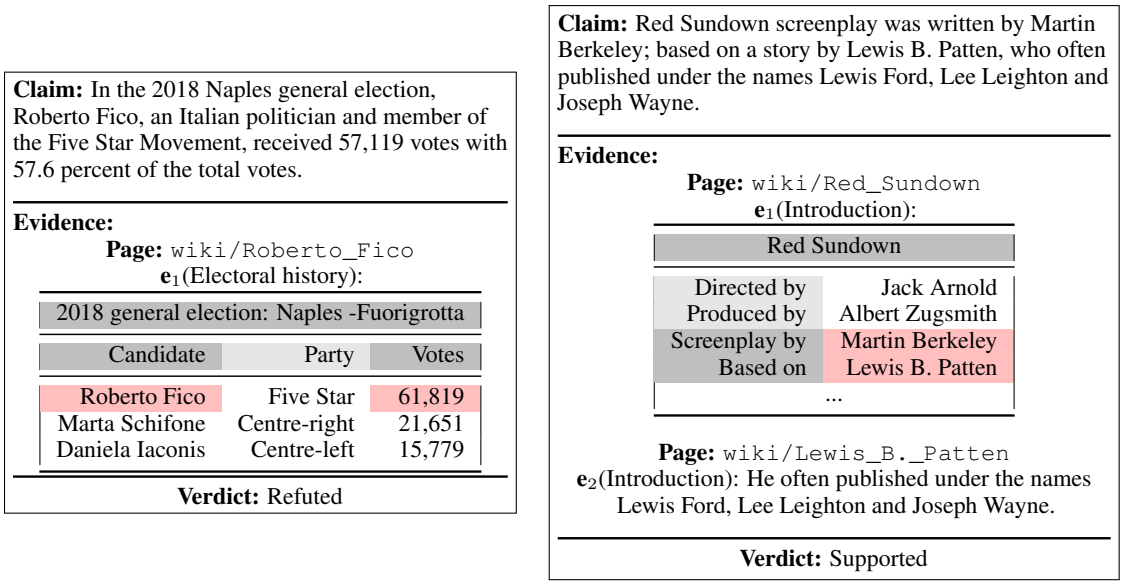

Figure 1: FEVEROUS sample instances. Evidence in tables is highlighted in red. Each piece of evidence e_i has associated context, i.e. page, section title(s) and the closest row/column headers (highlighted in dark gray). Left: evidence consists of two table cells refuting the claim. Right: Evidence consists of two table cells and one sentence from two different pages, supporting the claim.

2 Task Description

Given a human-authored claim, systems had to first retrieve evidence from Wikipedia in the form of sentences and table cells, each accompanied by the page/section titles and column headers they were found under respectively. They then had to classify whether the claim is SUPPORTED or REFUTED based on the evidence, or NOTENOUGHINFO if the claim cannot be verified. System responses would be scored both on the evidence retrieval and the label classification. Note that unlike in the original FEVER shared task, (partial) evidence needs to be provided for the NOTENOUGHINFO claims. Each claim in the FEVEROUS dataset could have multiple ways of being verified, which is represented in the different *evidence sets* - each with potentially multiple pieces of evidence. The participating systems only had to provide one complete evidence set for their response to be considered correct.

2.1 Dataset

We provided the training and development datasets through the FEVER website[1] and as an open source dataset[2]. A reserved portion of the dataset was released as a blind test set without the gold annotations (labels + evidence) to be used in the final phase of the challenge. The training data and the blind test set are described in (Aly et al., 2021), with each split's label distribution being thus known in advance. The label distribution of the dataset is only roughly balanced for the blind test set. The number of evidence sets with only textual evidence is slightly higher than sets that contain only textual evidence or sets that require a combination of different evidence types (c.f. Table 1).

	Train	Dev	Test
Supported	41,835	3,908	3,372
Refuted	27,215	3,481	2,973
NEI	2,241	501	1,500
Total	71,291	7,890	7,845
$E_{Sentences}$	31,607	3,745	3,589
E_{Cells}	25,020	2,738	2816
$E_{Sentences+Cells}$	20,865	2,468	2062

Table 1: Quantitative characteristics in each split of FEVEROUS, with $E_{Sentences}$, E_{Cells}, and $E_{Sentences+Cells}$ being claims requiring only sentence evidence, cell evidence, or both, respectively.

[1] https://fever.ai/dataset/feverous.html

[2] https://doi.org/10.5281/zenodo.4911507

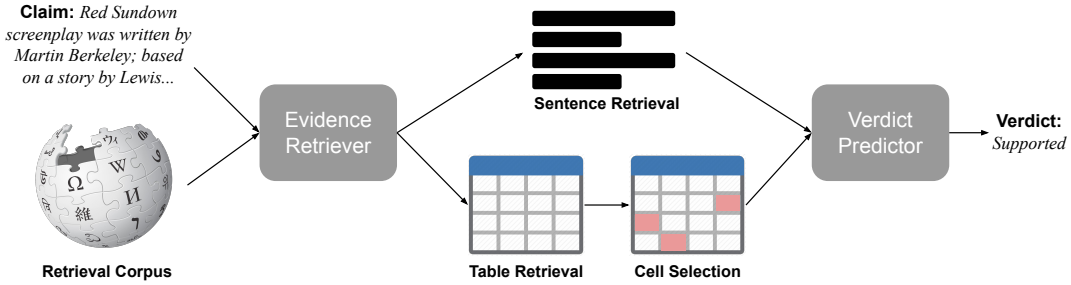

Figure 2: The pipeline of the FEVEROUS baseline, illustration taken from (Aly et al., 2021).

2.2 Submissions

The FEVEROUS shared task was hosted as a challenge on EvalAI[3] where participants were invited to submit predictions against the blind test set. Participants had about three days (24th July to 27th July 2021) starting from the release of the unlabeled blind test set to submit up to three submissions. Only a team's final submission was considered. The platform was open to submission on the development split one month prior, to allow participants to become familiar with the submission environment.

2.3 Baseline

The FEVEROUS baseline, shown schematically in Figure 2, employs a multi-stage retrieval pipeline, followed by a verdict prediction module. The most relevant documents are retrieved using a combination of entity matching and TF-IDF. The latter is then used to to extract the most relevant sentences and tables from the selected documents. To retrieve relevant cells from extracted tables, a cell extraction model linearizes them and treats the extraction as a sequence labelling task. A RoBERTa classifier (Liu et al., 2019) pre-trained on multiple NLI datasets, and fine-tuned on the FEVEROUS data[4], then predicts the veracity of the claim using the retrieved evidence and its context. Since the FEVEROUS dataset is imbalanced regarding NEI labels (5% of claims), the baseline additionally samples artificial NEI instances for training by partially removing evidence pieces from annotations. This baseline substantially outperforms the sentence-only and table-only baselines (Aly et al., 2021).

[3] https://eval.ai/web/challenges/challenge-page/1091

[4] For early reported scores of the baseline the classifier was not trained on the entire FEVEROUS data. These preliminary scores were almost identical to the final ones, with accuracy, evidence precision, and F_1 being around 0.01 points higher than here, but FEVEROUS score being marginally lower.

2.4 FEVEROUS score

Similar to the previous shared tasks of the FEVER Workshop (Thorne et al., 2018b), the scoring in FEVEROUS considers both the evidence retrieval and the claim labels. While we track the scores for each of these aspects individually (label accuracy, evidence P/R/F1), we use the FEVEROUS score defined in Aly et al. (2021) as the primary score for the challenge, defined as follows. For a given claim, a prediction is considered correct only if at least one complete gold evidence set E is a subset of the predicted evidence $\hat{E}$ and the predicted label is correct. We recognise that the evidence annotations are unlikely to be exhaustive, and measuring precision would penalise correct evidence missed by the annotators. Instead, we set a limit to the number of evidence pieces systems can return for each claim and allow only 5 predicted sentences items and 25 predicted table cells (the latter include table captions and list items). If additional evidence was returned it was discarded without penalty.

3 Results

The results for all submissions to the shared task are in Table 2. Further break downs are provided with results of each class (SUPPORTED, REFUTED NOTENOUGHINFO) in Table 2, and according to different types of evidence needed (textual-only, tabular-only, or both) in Table 4. The latter results are further analysed in Table 6 for each type of evidence. In Table 3, we further report results for claims that are particularly challenging, requiring numerical reasoning, multi-hop reasoning, entity disambiguation, or search terms beyond entities mentioned in the claim. Each of the 12 teams that submitted results for the blind test phase was invited to submit a paper to be reviewed in the FEVER workshop, as well as a 200-word descrip-

Rank	Team Name	Label Accuracy	Evidence Precision	Evidence Recall	Evidence F1	FEVEROUS Score
1	Bust a move!	0.56	0.08	**0.43**	0.13	**0.27**
2	Papelo	**0.58**	0.07	0.35	0.12	0.26
3	NCU	0.52	0.1	0.39	0.16	0.25
4	Z team	0.49	0.08	**0.43**	0.13	0.23
5	EURECOM_Fever	0.48	**0.14**	0.34	**0.2**	0.2
6	*Baseline*	0.46	0.10	0.29	0.15	0.18
7	Saturday Night Fever	0.48	0.11	0.29	0.16	0.18
8	Martin Funkquist	0.43	0.06	0.28	0.1	0.13
9	Albatross	0.41	0.06	0.19	0.1	0.12
10	METUIS	0.39	0.05	0.1	0.06	0.06
11	ChaCha	0.42	0.03	0.1	0.04	0.04
12	seda kaist	0.43	0.07	0.12	0.09	0.05
13	qmul uou iiith	0.4	0.02	0.03	0.03	0.02

Table 2: Final results of the blind test phase of the FEVEROUS challenge.

tion of their approach. 7 teams sent us system descriptions, six of which also submitted a paper. The descriptions appear in Appendix A (as sent by the authors except from minor typographic corrections), with the accompanying paper citation if one was submitted. In the remainder of this section we present our observations on the techniques used by the participants. The architecture followed by participating teams consisted of evidence retrieval followed by verdict prediction. Evidence retrieval was decomposed into page retrieval, followed by selecting textual (i.e. sentences) and tabular evidence (i.e. table cells) from the retrieved pages. Verdict prediction combined the retrieved evidence to return a label for the claim.

Page retrieval Page retrieval was mostly kept simple relying on term-matching for efficiency *Bust a move!*, *NCU* and *METUIS* used BM25 (the latter two using the implementation of Anserini (Yang et al., 2017)), while *Martin Funkquist* used vanilla TF-IDF matching. *Papelo* and *Albatross*, following the baseline, combined TF-IDF matching with entity matching, and *EURECOM_Fever* reranked its results with a BERT model pre-trained on MS-MARCO (Nguyen et al., 2016). Overall, focusing on the entities in the claim for page retrieval was found to be beneficial by the participants.

Sentence selection In order to select sentences to used as evidence, many teams used continuous representations in order to capture semantic affinity with the claim. *Bust a move!* applied a three stage-process to the task consisting of multi-hop dense passage-retrieval (Xiong et al., 2021) trained on FEVEROUS data, followed by BM25 filtering to ensure that sentences containing named entities mentioned in the claim are not ranked below those sentences semantically related to the claim but for different entities, and a final re-ranking step using a fine-tuned RoBERTa model. The latter is trained iteratively using a scheme to identify hard negative examples using previous versions of the model. Their method performs better than other systems on claims where the disambiguation of a claim's entity was a major challenge or when an article's name is not mentioned in the claim itself (c.f. Table 3), suggesting that their negative sampling method is effective. Papelo used a fine-tuned RoBERTa model for sentence selection combined with a next hop predictor based on T5 (Raffel et al., 2020) that aims to retrieve evidence complementing the pieces already retrieved. *NCU* used a BERT evidence classifier fine-tuned on FEVEROUS data, while *METUIS* developed a BERT-based QA model using the data provided. *Martin Funkquist* and *EURECOM_Fever* used TF-IDF matching following the baseline.

Table/cell selection Table selection was often done term-based using the same approaches as for page selection, i.e. TF-IDF as in the baseline (*Papelo, EURECOM_Fever*) and BM25 (*Bust a move!*), while *Martin Funkquist* used the dense table retriever of Herzig et al. (2021). *NCU* considers all tables in retrieved documents for cell extrac-

Team Name	Numerical Reasoning	Multi-hop Reasoning	Entity Disambiguation	Search terms not in Claim
Bust a move!	0.11	**0.14**	**0.20**	**0.15**
Papelo	**0.21**	0.10	0.10	0.02
NCU	0.10	**0.14**	**0.20**	0.12
Z team	0.10	0.13	0.18	0.11
EURECOM_Fever	0.07	0.12	0.17	0.10
Baseline	0.07	0.11	0.12	0.12
Saturday_Night_Fever	0.07	0.12	0.12	0.11
Martin Funkquist	0.01	**0.14**	0.12	0.09
Albatross	0.02	0.09	0.10	0.12
METUIS	0.04	0.00	0.04	0.01
ChaCha	0.03	0.01	0.01	0.00
seda_kaist	0.02	0.01	0.01	0.00
qmul_uou_iiith	0.02	0.01	0.01	0.01

Table 3: FEVEROUS scores on the blind test phase, requiring numerical reasoning (740 samples), multi-hop reasoning (1,195 samples), entity disambiguation (200 samples) and search terms beyond entities mentioned in claim (193 samples).

tion. The cells from the tables were often chosen using the same approach used by teams to select sentences from documents but trained on tabular data from the task (*Bust a move!, Papelo, NCU, EURECOM_Fever*). Cells are treated as text by linearzing them through concatenating their content and context with special markup. The teams considered as context a table's caption, table headers, and the page name, with the latter improving scores substantially for *NCU*. While *Bust a move!* and *Papelo* train a separate model for sentence and cell retrieval, *NCU* trains a single model on the joint tabular and textual data. *Martin Funkquist* and *METUIS* used the TAPAS QA model (Herzig et al., 2020) for retrieving cells. Using continuous representations to retrieve sentences and cells from retrieved documents have generally been successful, however, using specialised methods (i.e. TAPAS) explored by participants for table retrieval and cell selection seems to be have been less successful. Instead, term-based table retrieval, and treating tables as sequences of cells was overall more successful.

Evidence retrieval from different locations Claims requiring information from two or more different sections or articles (termed multi-hop reasoning in FEVEROUS), was a challenge to all systems, with neither of the two systems employing multihop evidence retrieval (*Bust a move!*, *Papelo*) scoring better (c.f. Table 3). However, we note that both teams' multihop evidence retrieval focus on the iterative retrieval of evidence, while for multihop claims labelled in FEVEROUS direct semantic matching with only the claim is sufficient in many cases.

Verdict prediction For verdict prediction, the top two teams developed models taking into account the fact that the evidence during testing will be noisy given that retrieval is imperfect. Thus they trained models using retrieved evidence (*Papelo*) or combining it with gold evidence from the data (*Bust a move!*). The latter considered all evidence to be of tabular form and used two instantiations of a TAPAS model, whose predictions were aggregated using an MLP. On the other hand, *Papelo* considered all evidence to be of sentence form using a simple markup to encode the table structure, and trained a T5 model on FEVEROUS data. In addition they facilitated the handling of mathematical reasoning at this stage by encoding numbers and relations between into "math hints" that were added as a prefix to the input. By using these hints, *Papelo* achieves by far the highest scores on claims that require numerical reasoning, as seen in Table 3. As this type of reasoning is typically also more relevant to claims requiring tabular evidence, it is possibly part of the reason *Papelo* performs substantially better in such claims (Table 4). Yet, all systems appear to struggle when a claim requires both textual and tabular evidence, as seen in Table 4. With exception to *Papelo*, the score tends to follow the tabular-only performance, which was more challenging for most systems. Following the baseline, *NCU*, *EURECOM_Fever* and *Albatross* treated all evidence as text and relied on some form of pretrained NLI model fine-tuned to the data from the shared task. *Albatross* fine-tuned several existing NLI models and used majority voting to obtain

the final verdict. METUIS used a pre-trained NLI model without further tuning which was applied to each piece of evidence retrieved, combining the predictions heuristically. Finally, *Martin Funkquist* handled textual evidence using RoBERTa and tabular evidence using TAPAS, aggregating their results using an MLP.

Team Name	Textual	Tabular	Combined
Bust a move!	**0.35**	0.23	0.22
Papelo	0.28	**0.33**	0.20
NCU	0.32	0.21	**0.23**
Z team	0.29	0.16	0.22
EURECOM_Fever	0.29	0.14	0.15
Baseline	0.27	0.12	0.12
Saturday_Night_Fever	0.26	0.11	0.14
Martin Funkquist	0.22	0.04	0.09
Albatross	0.26	0.03	0.01
METUIS	0.12	0.03	0.01
ChaCha	0.05	0.05	0.03
seda_kaist	0.05	0.05	0.03
qmul_uou_iiith	0.03	0.02	0.01

Table 4: FEVEROUS scores on the blind test phase, only considering samples that require exclusively textual evidence, tabular evidence, or both, respectively.

4 Analysis

Handling NOTENOUGHINFO (NEI) claims in FEVEROUS is substantially more challenging than in the FEVER shared task for two reasons: i) (partial) evidence must be retrieved for a prediction to be considered correct, ii) the dataset is very imbalanced, with relatively few NEI instances in the training set. As seen in Table 5, NEI performance was poor on the whole, with the top two systems opting not to predicting any NEI instances. In contrast to *Bust a move!*, *Papelo* by design does not predict any NEI instances, replacing instances in the training set labelled as NEI with Supported, turning the task into a binary classification task. While *Papelo* explored sampling artificial NEI instances, by labelling any instance with incomplete extracted evidence as NEI, their model performs much worse in this scenario, overpredicting NEI. A possible cause is that their prediction model is still trained on noisy evidence, creating the additional challenge of distinguishing a complete evidence set with possibly irrelevant evidence from an incomplete evidence set. This further presents an explanation for *Bust a move!*'s performance on NEI, as they also train their prediction model on both complete and incomplete evidence, which makes it more robust to imperfect retrieval for supported and refuted instances, yet making it impossible for the model to correctly distinguish these from instances with not enough information. Interestingly, worse overall systems did better on NEI predictions, with *Albatross* and *METUIS* receiving a relatively balanced F_1 score across all classes. This can possibly be attributed to their explicit treatment of the NEI class, with *METUIS* using a verdict heuristic to predict the NEI class if none of the extracted evidence pieces provides enough confidence in supporting or refuting a claim. They also report that their model overpredicts NEI instances on the development split, suggesting that it should be fine-tuned on the dataset.

Team Name	Supported	Refuted	NEI
Bust a move!	**0.66**	0.57	0.00
Papelo	0.65	**0.63**	0.00
NCU	0.61	0.54	0.26
Z team	0.62	0.38	0.01
EURECOM_Fever	0.58	0.46	0.19
Baseline	0.55	0.47	0.26
Saturday_Night_Fever	0.58	0.48	0.17
Martin Funkquist	0.60	0.01	0.00
Albatross	0.46	0.45	**0.29**
METUIS	0.35	0.49	**0.29**
ChaCha	0.53	0.35	0.11
seda_kaist	0.52	0.36	0.09
qmul_uou_iiith	0.32	0.51	0.17

Table 5: Label F_1 score per-class.

We measured how the metric of the FEVEROUS shared task correlates to its performance on both components of the task, namely evidence retrieval and veracity prediction and how the FEVEROUS scores compare to the scores obtained in FEVER (Thorne et al., 2018b). As seen in Figure 3, both FEVER and FEVEROUS scores for systems participating in the respective shared tasks strongly positively correlate with an increased label accuracy (Pearson correlation of $\rho = 0.92$ for both). However, concerning the retrieval component, it can be seen that the FEVEROUS scores correlate with the evidence F1 ($\rho = 0.83$) more strongly than the FEVER scores ($\rho = 0.41$), especially in terms of recall ($\rho = 0.97$ and $\rho = 0.53$, respectively). Again, this is a consequence of correct NEI predictions not requiring any evidence in FEVER.

The FEVEROUS baseline system is stronger than the one proposed for FEVER relatively to the respective submitted systems, achieving a higher

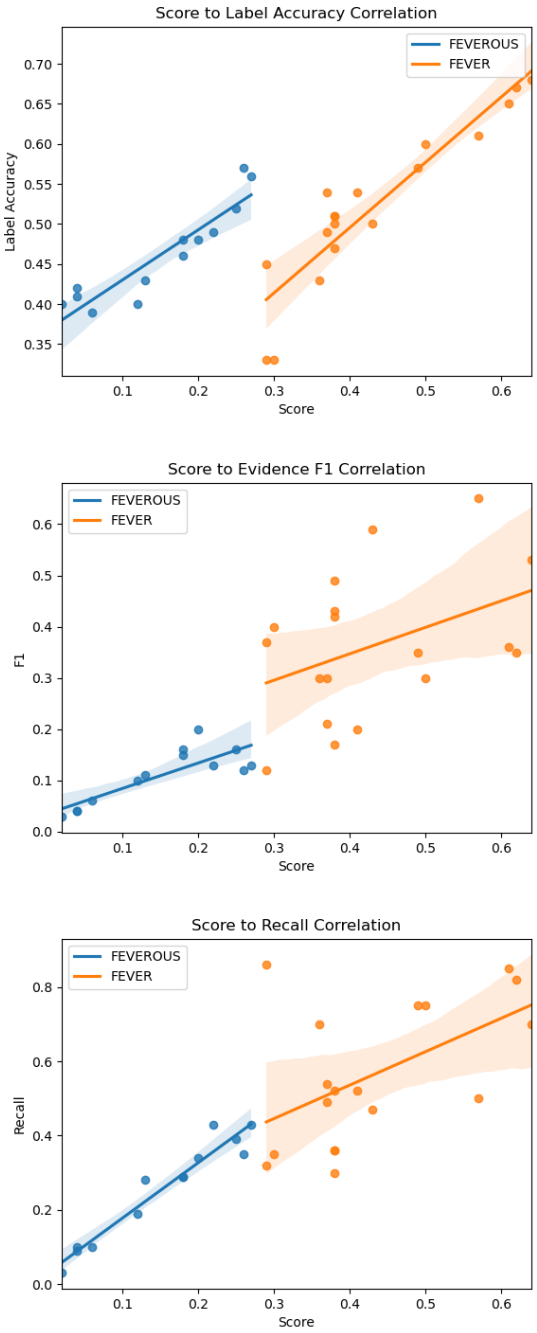

Figure 3: Correlation between FEVER/FEVEROUS score and Label Accuracy, F1, and Recall, respectively.

score than more than half of participating teams, while the FEVER baseline only performed better than around a fifth of all systems. Generally, FEVEROUS scores are generally much lower than FEVER scores, with retrieval scores being particularly low for FEVEROUS. While this is likely partly due to the NEIs being easier to predict and more numerous in FEVER (by only predicting NEI, a system would already get a score of 0.33), it might also be a result of artefacts in the original FEVER dataset, as a claim-only baseline was able to get an accuracy score of about 62% (Schuster et al., 2019), compared to a majority-class baseline of 33%. In contrast, the claim-only baseline on FEVEROUS achieves a score of 58% against a majority-class baseline of 56%. Yet, peculiarities of the FEVEROUS dataset observed by participants is a higher number of redundant evidence pieces than in FEVER (Malon, 2021) (likely a result of the higher complexity of evidence annotation), as well as a considerable number of claims that are refuted/NEI due to a single piece of information being incorrect/missing in an otherwise supported claim (Bouziane et al., 2021). Since claims are much longer in FEVEROUS than FEVER, such cases are much harder to identify. Similar to the FEVER 2.0 challenge (Thorne et al., 2019) where in a *build-it*, *break-it*, *fix-it* competition teams created adversarial attacks (*breakers*) against systems that were trained on the FEVER dataset (*builders*), to identify biases and weaknesses and address them (*fixers*), such a challenge might provide highly valuable insights to the FEVEROUS dataset to foster further research on this task.

Acknowledgements

We would like to thank Amazon for sponsoring the dataset generation and supporting the FEVER workshop and the FEVEROUS shared task. Rami Aly is supported by the Engineering and Physical Sciences Research Council Doctoral Training Partnership (EPSRC). James Thorne is supported by an Amazon Alexa Graduate Research Fellowship. Zhijiang Guo, Michael Schlichtkrull and Andreas Vlachos are supported by the ERC grant AVeriTeC (GA 865958).

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. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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. 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 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Mostafa Bouziane, Hugo Perrin, Amine Sadq, Thanh Nguyen, Aurélien Cluzeau, and Julien Mardas. 2021. FaBULOUS: Fact-checking based on understanding of language over unstructured and structured information. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 31 – 40. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *ICLR*, pages 1–14.

Martin Funkquist. 2021. Combining sentence and table evidence to predict veracity of factual claims using TaPaS and RoBERTa. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 92 – 101. Association for Computational Linguistics.

In-Zu Gi, Ting-Yu Fang, and Richard Tzong-Han Tsai. 2021. Verdict inference with claim and retrieved elements using roberta. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 60 – 66. Association for Computational Linguistics.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. 2020. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

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.

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.

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*.

Christopher Malon. 2021. Team Papelo at feverous: Multi-hop evidence pursuit. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 40–46. Association for Computational Linguistics.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1291–1299, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, and Thomas Mandl. 2021. The CLEF-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II*, volume 12657 of *Lecture Notes in Computer Science*, pages 639–649. Springer.

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. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring

the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Mohammed Adel Saeed, Giulio Alfarano, Khai Nguyen, Duc Pham, Raphael Troncy, and Paolo Papotti. 2021. Neural re-rankers for evidence retrieval in the FEVEROUS task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 108 – 113. Association for Computational Linguistics.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Orkun Temiz, Özgün Ozan Kılıç, Arif Ozan Kızıldağ, and Tuğba Taşkaya Temizel. 2021. A fact checking and verification system for FEVEROUS using a zero-shot learning approach. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 113 – 121. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. 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.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Wenhan Xiong, Xiang Li, Srini Iyer, Jingfei Du, Patrick Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021. Answering complex open-domain questions with multi-hop dense retrieval. In *International Conference on Learning Representations*.

Peilin Yang, Hui Fang, and Jimmy Lin. 2017. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1253–1256, New York, NY, USA. Association for Computing Machinery.

A System Description Summaries Submitted by Participants

A.1 Bust a move! (Bouziane et al., 2021)

We proposed a novel architecture to handle the joint retrieval and entailment over unstructured and structured information. To verify a claim, we first retrieve documents and filter the most relevant tables using BM25. Therefrom, our passage retriever extracts the relevant pieces of evidence, which can be either sentences or table cells. Finally, we obtain the verdict prediction by performing entailment using a TAPAS-based ensemble model. For retrieval, we proposed a novel training paradigm, Reinforced Adaptive Retrieval Embedding (RARE), which is inspired by reinforcement learning. It consists of re-ranking the BM25 retrieved hard-negative samples based on a snapshot of the embedding model of the last epoch. RARE samples better hard negatives, helping the model correct itself and preventing overfitting. For entailment, we proposed Noisy Entailment through Adapted Training (NEAT) that consists of two models trained on golden and noisy evidence sets, respectively. Together, they will see both relevant and irrelevant passages during training to make the ensemble more robust to noisy inputs at inference.

A.2 Papelo (Malon, 2021)

We develop a system for the FEVEROUS fact extraction and verification task that ranks an initial set of potential evidence and then pursues missing evidence in subsequent hops by trying to generate it, with a "next hop prediction module" whose output is matched against page elements in a predicted article. Seeking evidence with the next hop prediction module continues to improve FEVEROUS score for up to seven hops. Label classification is trained on possibly incomplete extracted evidence chains, utilizing hints that facilitate numerical comparison. The system achieves .281 FEVEROUS score and .658 label accuracy on the development set, and finishes in second place with .259 FEVEROUS score and .576 label accuracy on the test set.

A.3 NCU (Gi et al., 2021)

Our 3rd place FEVEROUS system is a three-stage model consisting of document retrieval, element retrieval, and verdict inference. Our document retrieval utilizes Anserini (Yang et al., 2017), an information retrieval toolkit built on Lucene. For element retrieval, we experiment two different approaches, the Anserini and the BERT model, to select relevant elements from documents retrieved in previous stage. For the third stage, we adopt the RoBERTa NLI model pre-trained on well-known NLI datasets, including SNLI, MNLI, FEVER-NLI, ANLI (Nie et al., 2020), and experiment on its variants with aggregation method to fully utilize the semantic information of retrieved elements in previous stage. Our system improves the FEVEROUS baseline in two aspects. First, while the baseline retriever pays attention to literally relevance with a combination method of entity matching and TF-IDF, we fine-tune the BERT model to integrate more semantic relevance for finding evidential elements and downstream verdict inference. Second, the baseline predictor uses the concatenation of claim and elements as input, having a constraint of maximum length. We experiment several ways to include more elements for verdict inference. These improvements allow us to achieve 0.29 feverous score on the development set and 0.25 feverous score on the blind test set, both outperforming the FEVEROUS baseline.

A.4 EURECOM_Fever (Saeed et al., 2021)

It is clear that enhancing evidence retrieval plays a vital role in any fact-checking system. In our submission, 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., 2017). We extend the baseline by providing a two-stage re-ranking process in the spirit of simple IR systems: (a) first, numerous pages to a given query are retrieved from a corpus using entity-matching and TF-IDF (Chen et al., 2017) and (b) second, the pages are scored and re-ranked using a more computationally-demanding method. Given that neural ranking methods have shown success in the IR community (Guo et al., 2020), we used one as part of our extension to the baseline, where a re-ranker provides a score for every (query, table) pair. We then retain the tables with the top scores. The re-ranker is based on a pre-trained BERT model that is fine-tuned on the passage re-ranking task of the MS MACRO (Nguyen et al., 2016) dataset to minimize the binary cross-entropy loss. This extension to the baseline was enough to beat it, exhibiting that having higher recall with a computationally-demanding method would be more effective for evidence retrieval than standard mechanisms.

A.5 Martin Funkquist (Funkquist, 2021)

The proposed system consists of three main parts: document retrieval, evidence retrieval and label prediction. The first part retrieves the most relevant documents using TF-IDF vector similarity scores between the claim and the title and body text of the documents. Then evidence is retrieved from these documents using similarity scores between TF-IDF vectors to retrieve the textual evidence and similarity scores between dense vectors created by fine-tuned TaPaS models to retrieve tabular evidence. Finally, the evidence is passed through a dense neural network to produce a veracity label, where the input is vectors created by a pre-trained RoBERTa for the sentence evidence and a TaPaS model for the table evidence.

A.6 Albatross

For the retrieval part, we experimented with Spacy NER based on Transformers and FastText matching instead of TFIDF. We also analyzed performance change across parameters like page count, table count, etc in the TFIDF module of baseline implementation. For verdict prediction, we fine tuned several publicly available NLI models on the competition data. We also tried a majority vote strategy

for creating the test predictions using the various verdict prediction models that we had trained.

A.7 METUIS (Temiz et al., 2021)

We propose a pipeline that retrieves documents by using Anserini indexing on top of the Wikipedia dump. After the document retrieval, evidence related to the claim is selected by using Bert-Large-Cased Question Answering model, and the results of the QA model are sorted by using Universal Sentence Encoder score, which measures the similarity between the claim and the document portion. The final verdict of the claim is determined by the XLNET natural language inference model, which compares the evidence and the claim. Other than the sentence evidence, the cell evidence is obtained by TAPAS Table Question Answering model and by looking at the match score between the entities of the claim and the cell values. The pipeline is fully unsupervised, and all the models used in the pipeline require no pretraining.

B Further result tables

Team Name	Label Accuracy	Evidence Precision	Evidence Recall	Evidence F1	FEVEROUS Score
Textual-only					
Bust a move	0.54	0.06	0.57	0.11	0.35
Papelo	0.60	0.06	0.35	0.10	0.28
NCU	0.56	0.11	0.48	0.17	0.32
Z team	0.47	0.06	0.57	0.11	0.29
EURECOM_Fever	0.51	0.15	0.49	0.23	0.29
Baseline	0.50	0.13	0.43	0.20	0.27
Saturday_Night_Fever	0.50	0.12	0.43	0.19	0.26
Martin Funkquist	0.40	0.07	0.49	0.13	0.22
Albatross	0.49	0.10	0.43	0.16	0.26
METUIS	0.44	0.05	0.17	0.08	0.12
ChaCha	0.39	0.02	0.13	0.03	0.05
seda_kaist	0.39	0.02	0.13	0.03	0.05
qmul_uou_iiith	0.41	0.02	0.05	0.03	0.03
Tabular-only					
Bust a move	0.58	0.07	0.36	0.12	0.23
Papelo	0.63	0.07	0.44	0.13	0.33
NCU	0.49	0.07	0.35	0.12	0.20
Z team	0.43	0.07	0.36	0.12	0.16
EURECOM_Fever	0.44	0.09	0.28	0.13	0.14
Baseline	0.47	0.07	0.22	0.11	0.12
Saturday_Night_Fever	0.46	0.07	0.22	0.11	0.11
Martin Funkquist	0.30	0.04	0.16	0.06	0.04
Albatross	0.34	0.01	0.04	0.01	0.02
METUIS	0.41	0.03	0.09	0.04	0.05
ChaCha	0.41	0.02	0.13	0.04	0.05
seda_kaist	0.42	0.02	0.13	0.04	0.05
qmul_uou_iiith	0.51	0.03	0.04	0.03	0.02
Combined					
Bust a move	0.59	0.11	0.30	0.16	0.22
Papelo	0.50	0.10	0.27	0.14	0.19
NCU	0.52	0.13	0.33	0.18	0.23
Z team	0.62	0.11	0.30	0.16	0.22
EURECOM_Fever	0.49	0.18	0.19	0.18	0.15
Baseline	0.46	0.15	0.16	0.16	0.13
Saturday_Night_Fever	0.49	0.15	0.17	0.16	0.14
Martin Funkquist	0.65	0.09	0.12	0.10	0.09
Albatross	0.34	0.08	0.01	0.02	0.01
METUIS	0.29	0.06	0.04	0.05	0.01
ChaCha	0.48	0.05	0.05	0.05	0.03
seda_kaist	0.46	0.04	0.03	0.04	0.02
qmul_uou_iiith	0.27	0.03	0.02	0.02	0.01

Table 6: Results of the blind test phase of the FEVEROUS challenge, only considering samples that require exclusively textual evidence (top), tabular evidence (middle), and both (bottom), respectively.

Team Name	Label Accuracy	Evidence Precision	Evidence Recall	Evidence F1	FEVEROUS Score
Numerical Reasoning (740)					
Bust a move	0.55	0.09	0.22	0.13	0.11
Papelo	0.62	0.11	0.32	0.16	0.21
NCU	0.46	0.10	0.20	0.13	0.10
Z team	0.41	0.09	0.22	0.13	0.08
EURECOM_Fever	0.42	0.09	0.20	0.13	0.07
Baseline	0.38	0.08	0.16	0.11	0.07
Saturday_Night_Fever	0.44	0.08	0.16	0.11	0.07
Martin Funkquist	0.32	0.04	0.10	0.06	0.01
Albatross	0.31	0.02	0.05	0.03	0.02
METUIS	0.33	0.03	0.06	0.04	0.04
ChaCha	0.41	0.02	0.07	0.03	0.03
seda_kaist	0.41	0.02	0.07	0.03	0.02
qmul_uou_iiith	0.51	0.03	0.03	0.03	0.02
Multi-hop Reasoning (1195)					
Bust a move	0.55	0.09	0.20	0.12	0.14
Papelo	0.48	0.07	0.13	0.09	0.10
NCU	0.46	0.12	0.21	0.15	0.14
Z team	0.59	0.09	0.20	0.12	0.13
EURECOM_Fever	0.47	0.18	0.16	0.17	0.12
Baseline	0.44	0.13	0.15	0.14	0.11
Saturday_Night_Fever	0.48	0.14	0.15	0.15	0.12
Martin Funkquist	0.65	0.09	0.18	0.12	0.14
Albatross	0.38	0.10	0.11	0.10	0.09
METUIS	0.32	0.05	0.01	0.02	0.00
ChaCha	0.47	0.03	0.03	0.03	0.01
seda_kaist	0.46	0.03	0.03	0.03	0.01
qmul_uou_iiith	0.23	0.03	0.01	0.02	0.01
Entity Disambiguation (200)					
Bust a move	0.40	0.07	0.34	0.11	0.20
Papelo	0.41	0.05	0.14	0.08	0.10
NCU	0.49	0.09	0.27	0.14	0.20
Z team	0.38	0.07	0.35	0.11	0.18
EURECOM_Fever	0.41	0.13	0.26	0.17	0.17
Baseline	0.42	0.10	0.21	0.14	0.12
Saturday_Night_Fever	0.40	0.12	0.22	0.15	0.12
Martin Funkquist	0.37	0.07	0.25	0.11	0.12
Albatross	0.43	0.08	0.17	0.11	0.10
METUIS	0.37	0.04	0.06	0.05	0.04
ChaCha	0.38	0.02	0.06	0.03	0.01
seda_kaist	0.37	0.02	0.05	0.03	0.01
qmul_uou_iiith	0.24	0.02	0.01	0.01	0.01
Search terms not in Claim (195)					
Bust a move	0.28	0.05	0.33	0.09	0.15
Papelo	0.26	0.02	0.09	0.03	0.02
NCU	0.38	0.07	0.24	0.11	0.12
Z team	0.26	0.05	0.35	0.09	0.11
EURECOM_Fever	0.33	0.08	0.23	0.12	0.10
Baseline	0.38	0.06	0.19	0.09	0.12
Saturday_Night_Fever	0.30	0.07	0.20	0.10	0.11
Martin Funkquist	0.27	0.05	0.25	0.08	0.09
Albatross	0.48	0.05	0.17	0.07	0.12
METUIS	0.37	0.02	0.03	0.02	0.01
ChaCha	0.27	0.01	0.05	0.01	0.00
sed_kaist	0.26	0.01	0.05	0.01	0.00
qmul_uou_iiith	0.22	0.01	0.01	0.01	0.01

Table 7: FEVEROUS scores on the blind test phase, grouped into their different challenges, with samples numbers in brackets.