QMUL-SDS at SCIVER: Step-by-Step Binary Classification for Scientific Claim Verification

Xia Zeng, Arkaitz Zubiaga
Queen Mary University of London

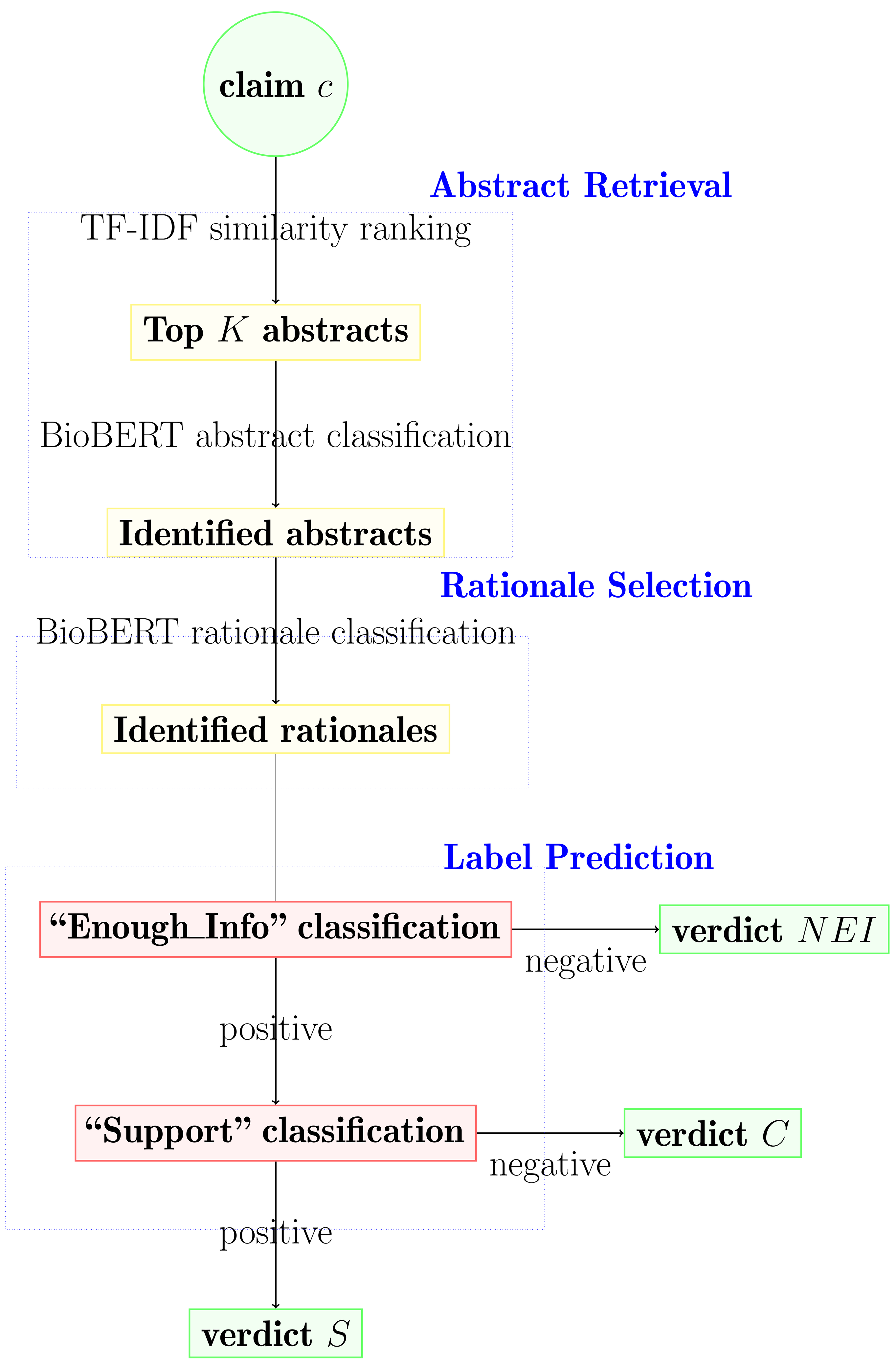
Introduction

- Scientific claim verification has gained increasing interest in the context of the ongoing COVID-19 pandemic.
- We propose an approach that performs scientific claim verification by doing binary classifications step-by-step.
- Our team was the No. 4 team on the leaderboard. We achieve substantial improvements over the baseline system without using extra data or increasing model size.

SCIVER Shared Task

- A benchmark scenario to test and compare scientific claim verification approaches.
- Given a scientific claim and a corpus of over 5000 abstracts, the task consists in (i) identifying abstracts relevant to the claim, (ii) delving into the abstracts to select evidence sentences relevant to the claim, and (iii) subsequently predicting claim veracity.
- **Evaluation.** Abstract-level evaluation and sentence-level evaluation.

Our System Overview



- **Overview of our step-by-step binary classification system.**
- NEI stands for "NOT_ENOUGH_INFO", C stands for "CONTRADICT" and S stands for "SUPPORT". Given claim c, our system first retrieves top K TF-IDF similarity abstracts out of the corpus, then uses a BioBERT binary classifier to further identify desired abstracts on top of that. With retrieved abstracts, our system then uses another BioBERT binary classifier to select rationales. We finally do label prediction in a two-step fashion, i.e. first make verdicts on "ENOUGH_INFO" or not and, if positive, then make verdicts on "SUPPORT" or not.

SCIFACT Dataset

- 1,409 expert-annotated biomedical claims.
- 5,183 abstracts from peer-reviewed publications.
- Each claim has a label out of *supports*, *refutes* and *not_enough_info*.

VERISCI Baseline System

- Abstract retrieval: TF-IDF similarity ranking -> TOP-K documents
- Rationale selection: RoBERTa-large with sigmoid function -> sentences whose relevance score is higher than threshold T
- Label prediction: RoBERTa-large classifier trained on FEVER and SCIFACT -> labels

Subtask Performance

Evidence Retrieval	Abstract Retrieval			Rationale Selection		
	P	R	F1	P	R	F1
Baseline	16.22	69.86	26.33	64.99	**70.49**	67.63
OurSystem	**62.75**	**74.16**	**67.98**	**77.08**	63.39	**69.57**
Label Prediction	**Oracle Evidence**			**Baseline Evidence**		
	P	R	F1	P	R	F1
Baseline	**90.75**	75.12	82.20	**56.42**	48.32	**52.06**
OurSystem	88.54	**81.33**	**84.78**	43.31	**52.63**	47.52

- Supervised training on abstract retrieval substantially reduces false positive predictions.
- Removing manual threshold on selection tasks simplifies the practice and has positive contributions.
- Two-step label prediction outperforms three-way label prediction on oracle evidence.
- **Improved label prediction module has worse performance with low-quality evidence inputs.**

Full Pipeline Results

Abstract-level	Label Only			Label+Rationale		
System	P	R	F1	P	R	F1
Baseline	47.51	47.30	47.40	46.61	46.40	46.50
OurSystem	**74.32**	**49.55**	**59.46**	**72.97**	**48.65**	**58.38**
Sentence-level	**Selection Only**			**Selection+Label**		
System	P	R	F1	P	R	F1
Baseline	44.99	47.30	46.11	38.56	40.54	39.53
OurSystem	**81.58**	**58.65**	**68.24**	**66.17**	**47.57**	**55.35**

- **Full pipeline performance on SCIFACT's test set**. Our system uses BioBERT-large for abstract retrieval and rationale selection and RoBERTa-large for two-step label prediction, all trained on SCIFACT train set and dev set.

Conclusions

- Concerning evidence retrieval, a classification based approach is better than a ranking based approach with manual thresholds.
- Two-step binary label prediction has better performance than three-way label prediction with limited training data.
- A more systematic design of automated fact-checking system is desired.

Acknowledgements: This work was supported by the Engineering and Physical Sciences Research Council (grant EP/V048597/1). Xia Zeng is funded by China Scholarship Council (CSC). This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045