NapierNLP at Touché: Fallacy Detection with Small Language Models and Model Fusion

Edinburgh Napier UNIVERSITY

Katarina Alexander PhD Student
Email: katarina.alexander@napier.ac.uk

Dr Dimitra Gkatzia and Dr Md Zia Ullah Supervisors
d.gkatzia, m.ullah@napier.ac.uk

Aims

- Subtask 1 of Touché Fallacy Detection task: Given an argument, determine whether it is fallacious[1,2,3] (it seems convincing, but does not lead to the presented conclusion)
- Utilise Small Language Models (SLMs) with under 5B parameters
- Experiment with two different ensemble types
- Test both using base data only, and using the enhanced data only
- 938 training examples (465 fallacies, 473 non)
- 233 testing examples (115 fallacy, 118 non)

id	informal_fallacies:332
text_base	Humans are omnivores & eating animals is natural. Get over it
argument_base	*claim: "*Eating animals is natural and people should accept it." *supports:* ["Humans are omnivores."]
text_enhanced	Humans are omnivores & eating animals is natural, and because it's natural, we should keep doing it. There's no reason to stop something that's part of our biology. Get over it.
argument_enhanced	*claim: "*Humans should continue eating animals" *supports:* ["Humans are omnivores, so eating animals is natural.", "Because eating animals is natural, it is justified to continue doing it.", "There is no reason to stop something that is part of our biology."]
fallacy_exists	1

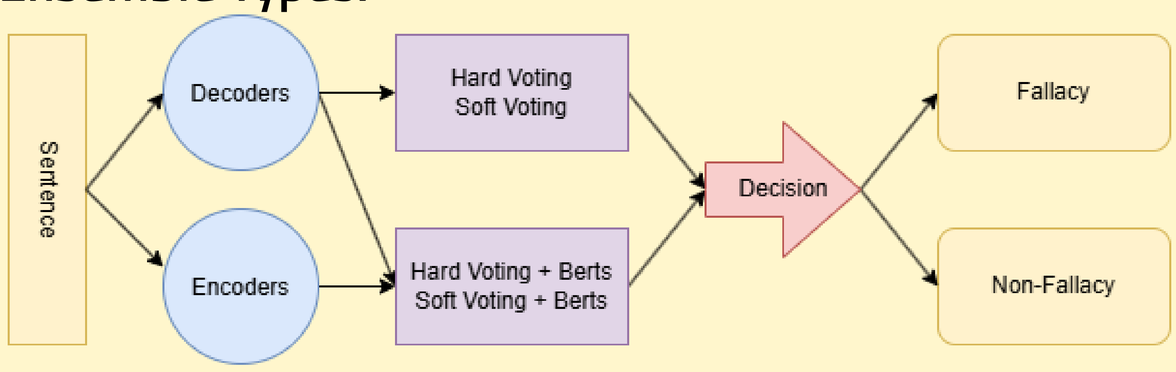
Example of a data entry used for this experiment

Small Language Models:

Training done using AutoModelForSequenceClassification on an 8GB memory GPU

Type:	Model:	Parameters:	Developer:	Released:
Decoder	GPT2	137M	OpenAI	2019
Decoder	GPT-Neo	1.37B	EleutherAI	2021
Decoder	Qwen2.5	1.54B	Alibaba Group	2024
Decoder	TinyLlama	1.1B	StatNLP Research Group	2024
Encoder	BERT		Google	2018
Encoder	RoBERTa		FacebookAI (META)	2019

Four Ensemble Types:



Hard Voting: Each model casts one predicted label. The majority label becomes the final classification.

Soft Voting: Each model outputs class confidence probabilities. The average is used for classification.

Base Data Results

- Came 7th with our SVB ensemble

Model	Precision	Recall	Accuracy	Macro F1
GPT2	.654	.722	.674	0.673
Qwen2.5	.730	**.774**	**.747**	**0.747**
GPT-Neo	**.742**	.574	.691	0.686
TinyLlama	.674	.757	.700	0.699
HV	.731	.757	.743	*0.743*
SV	.719	.757	.734	0.734
Bert	.602	.696	.622	0.621
Roberta	.636	**.774**	.670	0.667
HVB	.718	**.774**	.738	0.738
SVB	.712	**.774**	.734	0.734

Table 2: Performance of models on the test dataset: Trained and tested on the base data only

- Qwen2.5 was the best individual model for the base data — this would have placed 3rd in the competition
- Majority Vote ensembles (HV, HVB) outperformed the soft voting ensembles
- Adding BERT and RoBERTa did not improve model performance

- 53 of the test sentences were misclassified by every base ensemble, of these 24 were fallacies and 29 were non-fallacies.
- 2 sentences were correctly identified by all base ensembles and misidentified by all enhanced ensembles

Enhanced Data Results

- Came 7th with our SVB ensemble, 10th with HVB

Model	Precision	Recall	Accuracy	Macro F1
GPT2	.816	.887	.846	.845
Qwen2.5	**.926**	.870	.901	.901
GPT-Neo	.769	.896	.816	.815
TinyLlama	.906	**.922**	**.914**	**.914**
HV	.904	.896	.901	.901
SV	.897	.904	.901	.901
Bert	.835	.835	.837	.837
Roberta	.884	.861	.876	.876
HVB	.880	.896	.888	.888
SVB	.898	**.922**	.910	*.910*

Table 3: Performance of models on the test dataset: Trained and tested on the enhanced data only.

- TinyLlama was the best individual model for the enhanced data —this would have equalled the 6th place team
- Adding Bert and RoBERTa improved the soft voting ensemble, but reduced hard voting performance

- 16 of the test sentences were misclassified by every enhanced ensemble, of these 7 were fallacies and 9 were non-fallacies.
- 34 sentences were correctly identified by all enhanced ensembles, and incorrectly identified by all base ensembles

References

1. https://touche.webis.de/clef26/touche26-web/fallacy-detection.html
2. J. Kiesel, M. Feger, T. Hagen, S. Heineking, M. Heinrich, M. Fröbe, K. Boland, C. Mathews, W. Pertsch, J. Romberg, I. Zelch, S. Dietze, M. Hagen, M. Potthast, B. Stein, Overview of touché 2026 — argumentation systems, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, A. G. S. Herrera (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 17th International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.
3. J. Kiesel, M. Feger, T. Hagen, S. Heineking, M. Heinrich, M. Fröbe, K. Boland, C. Mathews, W. Pertsch, J. Romberg, I. Zelch, S. Dietze, M. Hagen, M. Potthast, B. Stein, Overview of touché 2026 — argumentation systems — extended version, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, 2026