Edinburgh Napier UNIVERSITY

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

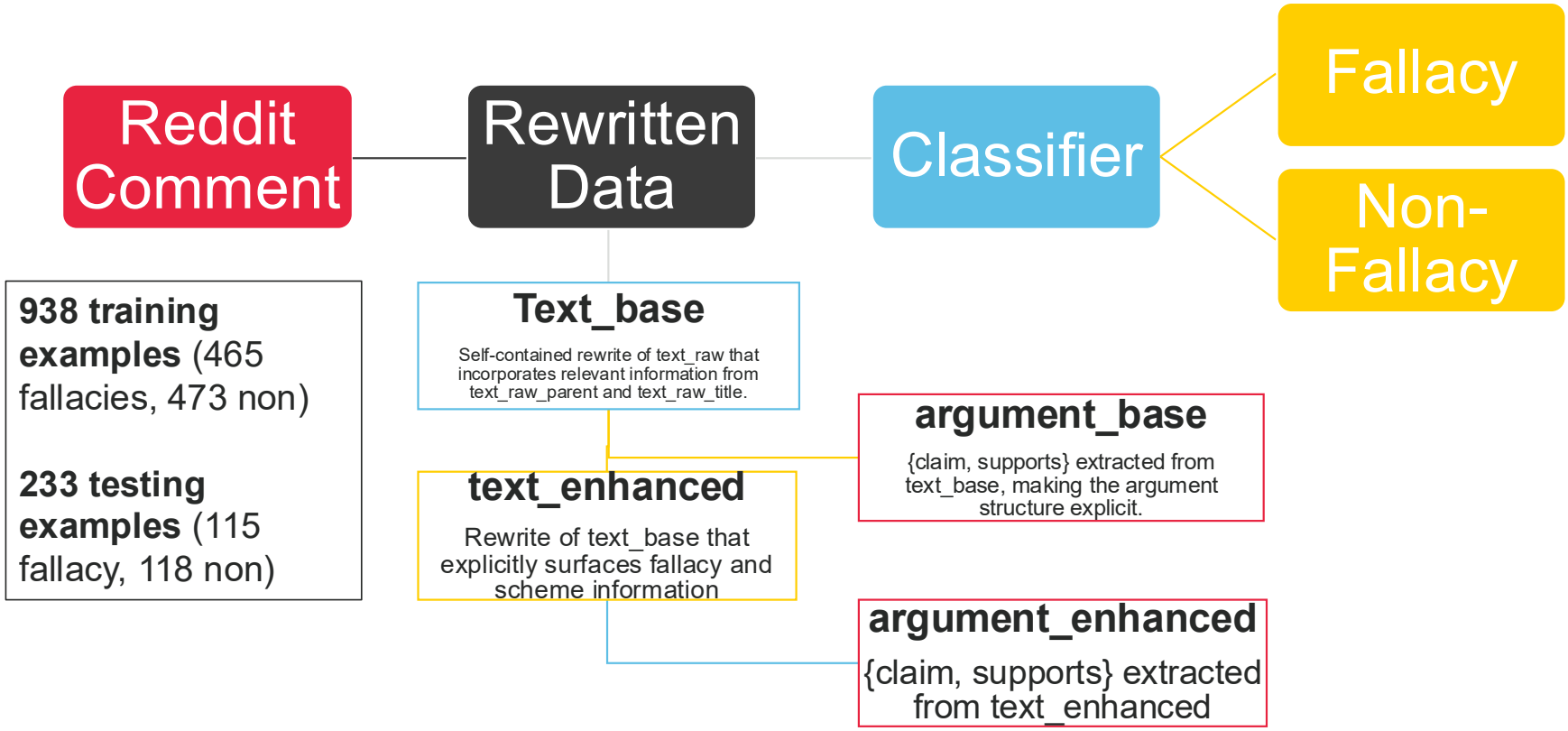
Katarina (Katja) Alexander – katarina.alexander@napier.ac.uk

Md Zia Ullah – M.Ullah@napier.ac.uk

Dimitra Gkatzia – D.Gkatzia@napier.ac.uk

Edinburgh Napier UNIVERSITY

Subtask 1: Fallacy Detection



Edinburgh Napier UNIVERSITY

Aim

- Utilise Small Language Models (SLMs) with under 5B parameters
  - These are small enough to run on commercially available GPUs so can be run locally on any GPU enabled laptop or computer (typically advertised for gaming)
- Experiment with two ensembles
  - Ensembles generally perform better as different models learn different patterns
- Experiment with adding encoder-only models to the ensembles
  - Adding in models with a different underlying structure can often improve ensembles as the differences are more pronounced

Edinburgh Napier UNIVERSITY

Models

SLM **GPT2**
137M Parameters

SLM **GPT-Neo 1.3B**
1.37B Parameters

SLM **Qwen2.5 1.5B**
1.54B Parameters

SLM **TinyLlama 1.1B**
1.1B Parameters

EM **BERT**
encoder-only

EM **RoBERTa**
encoder-only

1. openai-community/gpt2
2. EleutherAI/gpt-neo-1.3B
3. Qwen/Qwen2.5-1.5B
4. TinyLlama/TinyLlama_v1.1
5. Google/BERT
6. FacebookAI/RoBERTa

Edinburgh Napier UNIVERSITY

Ensembles

Hard Voting

- Each model casts one predicted label. The majority label across models becomes the final classification.

HV: 4 SLMs Only

HVB: 4 SLMs + BERT + RoBERTa

Soft Voting

- Each model outputs class confidence probabilities. Probabilities are averaged, and the final label follows the average.

SV: 4 SLMs Only

SVB: 4 SLMs + BERT + RoBERTa

Edinburgh Napier UNIVERSITY

Base Results

Model	Precision	Recall	Accuracy	Macro F1
GPT2	0.654	0.722	0.674	0.673
Qwen2.5	0.730	0.774	0.747	**0.747**
GPT-Neo	0.742	0.574	0.691	0.686
TinyLlama	0.674	0.757	0.700	0.699
SV	0.719	0.757	0.734	0.734
HV	0.731	0.757	0.743	**0.743**
Bert	0.602	0.696	0.622	0.621
Roberta	0.636	0.774	0.67	0.667
SVB	0.712	0.774	0.734	0.734
HVB	0.718	0.774	0.738	0.738

- We submitted our SVB run which came 7th in the competition
- Qwen2.5 was the best overall – alone it would have come 3rd
- Adding BERT and RoBERTa reduced the macro F1 of our hard voting ensemble

Edinburgh Napier UNIVERSITY

Enhanced Results

Model	Precision	Recall	Accuracy	Macro F1
GPT2	0.816	0.887	0.846	0.845
Qwen2.5	0.926	0.870	0.901	0.901
GPT-Neo	0.769	0.896	0.816	0.815
TinyLlama	0.906	0.922	0.914	**0.914**
SV	0.897	0.904	0.901	0.901
HV	0.904	0.896	0.901	0.901
Bert	0.835	0.835	0.837	0.837
Roberta	0.884	0.861	0.876	0.876
SVB	0.898	0.922	0.910	**0.910**
HVB	0.880	0.896	0.888	0.888

- We submitted our SVB and HVB ensembles
- TinyLlama was the best model overall, but would not have affected our placement
- Adding BERT and RoBERTa increased our soft voting ensemble, but decreased our hard voting ensemble

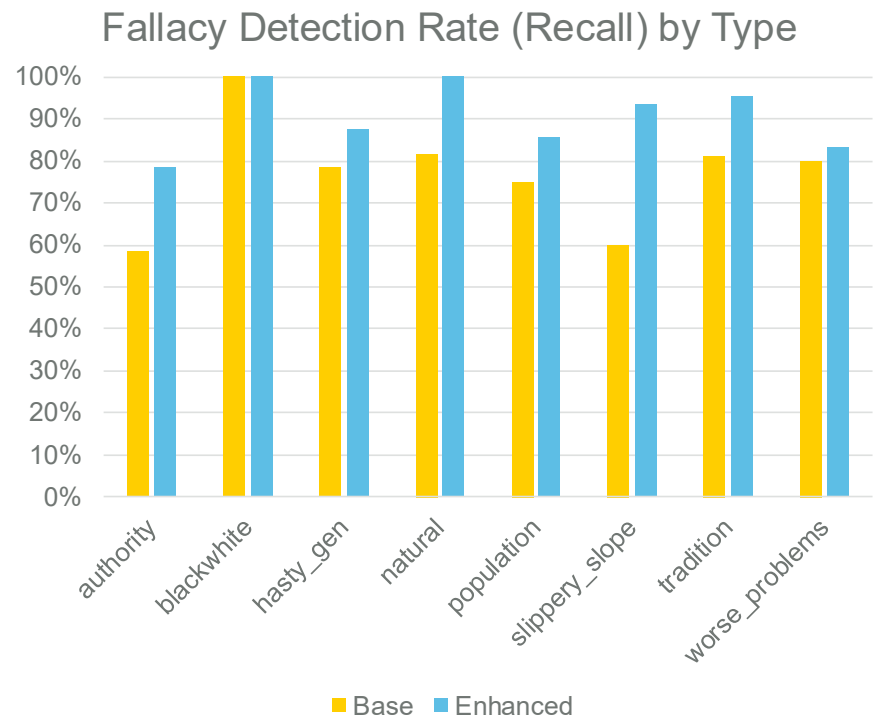
Failure Analysis

- 14 of the 233 sentences were misclassified by all eight ensemble methods,
- Of the correctly identified sentences, 155 of 233 were correctly identified by every single ensemble method, and 205 were correctly identified by every single enhanced data ensemble.
- Only two sentences were correctly identified by all base ensembles and misidentified by all enhanced ensembles: 401 (fallacy) and 1403 (non-fallacy).

Edinburgh Napier UNIVERSITY

Failure Analysis – Fallacy Type Recall

- Of the different fallacy types, none of the sentences tagged as black-or-white thinking were misclassified.
- The enhanced data ensembles also correctly identified every appeal to nature fallacy.
- The most common fallacy type to be misclassified was the appeal to authority.

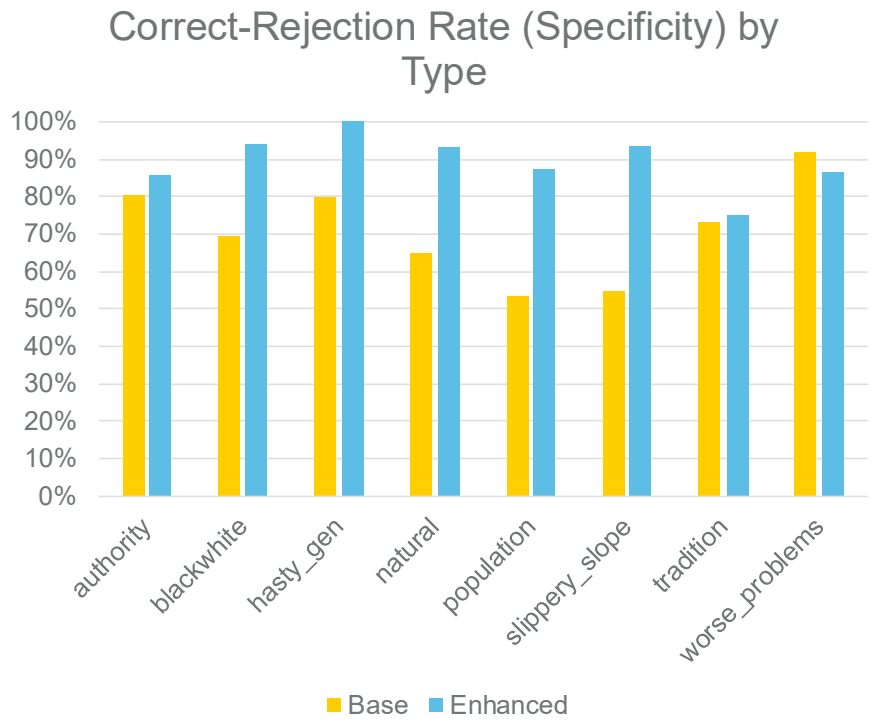


% correctly flagged as fallacy by Fallacy type

Edinburgh Napier UNIVERSITY

Failure Analysis – Fallacy Type Specificity

- The most common non-fallacy to be misclassified by the enhanced ensembles resembled the appeal to tradition fallacy
- By the base ensembles resembled the appeal to population
- The enhanced data ensembles correctly identified every non-fallacy resembling hasty generalization



% correctly identified as non-fallacy by fallacy type

Future Work

- Identifying other models which could be included in the ensembles, or removing models that may reduce the overall ensemble performance.
- Weighting to ensure models such as Qwen2.5 or TinyLlama have a larger impact on the overall ensemble.
- Re-training the models to access more of the data fields available, such as the raw text, parent and title may also have improved both the individual models and therefore the ensembles F1 scores.
- Adding prompting, either as an alternative to or in conjunction with our current method.

Thank you

Katja Alexander – katarina.alexander@napier.ac.uk

Md Zia Ullah – M.Ullah@napier.ac.uk

Dimitra Gkatzia – D.Gkatzia@napier.ac.uk