Detect human or machine text

Mihai-Octavian Ocnaru, Alexandru-Ionuț Mihalcea and Florin Brad *
University of Bucharest, Romania
*Bitdefender, Romania

mihai-octavian.ocnaru@s.unibuc.ro, alexandru-ionut.mihalcea@s.unibuc.ro, fbrad@bitdefender.com

1. Introduction

Machine-generated text detection has become increasingly important with the proliferation of powerful language models like ChatGPT, GPT-4, and other LLMs. The challenge of distinguishing between human-written and machine-generated text is significant, particularly when considering cross-domain generalization.

While previous studies have focused on larger models, we investigate the efficacy of more lightweight BERT variants (Tiny-BERT, Small-BERT, and Mini-BERT) for this task, with particular attention to their cross-domain generalization capabilities. Specifically, we:

- Evaluate three lightweight BERT variants on five domains from the M4 dataset
- Analyze cross-domain performance when models are trained on one domain and tested on others
- Compare the generalization capabilities of these smaller models against each other

Our findings provide insights into the trade-off between model size and performance in machine-generated text detection, which is valuable for applications with computational constraints.

2. Dataset and Preprocessing

We utilize the M4 dataset from Wang et al. (2023), following their exact methodology and focusing on five English domains: (1) **Wikipedia** articles, (2) **Reddit** ELI5 question-answer pairs, (3) **WikiHow** instructional guides, (4) **arXiv** scientific abstracts, and (5) **PeerRead** academic reviews.

The M4 dataset provides paired human-written and text generated by different large language models: **DaVinci**, **ChatGPT**, **Cohere**, **Dolly-V2**, **Bloomz**, **Flan-t5**, **Llama**. Our preprocessing adhered strictly to the M4 methodology, maintaining the minimum text length of 1000 characters and balanced train/validation/test splits (70/15/15%).

Dataset Statistics:

- **Human texts:** 1,528 avg. chars, 252,244 unique tokens
- **Machine texts:** 1,246 avg. chars, 275,455 unique tokens

While maintaining strict adherence to the original M4 dataset configuration, our research focuses specifically on evaluating lightweight BERT variants for cross-domain generalization assessment.

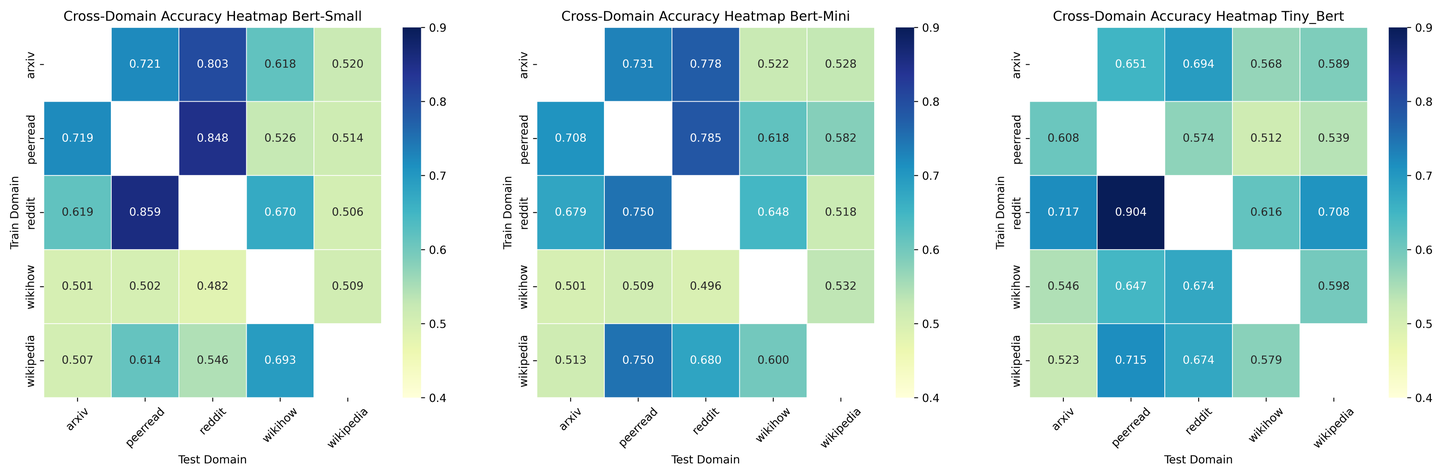
3. Models

We investigated three lightweight BERT variants: **Tiny-BERT:** (2 layers, 128 hidden size), **Tiny-BERT:** (4 layers, 256 hidden size), **Small-BERT:** (4 layers, 512 hidden size).

The models were fine-tuned using the following hyperparameters: learning rate: 2e-5, batch size: 32, training epochs: 5, weight decay: 0.05, maximum sequence length: 512, and AdamW optimizer.

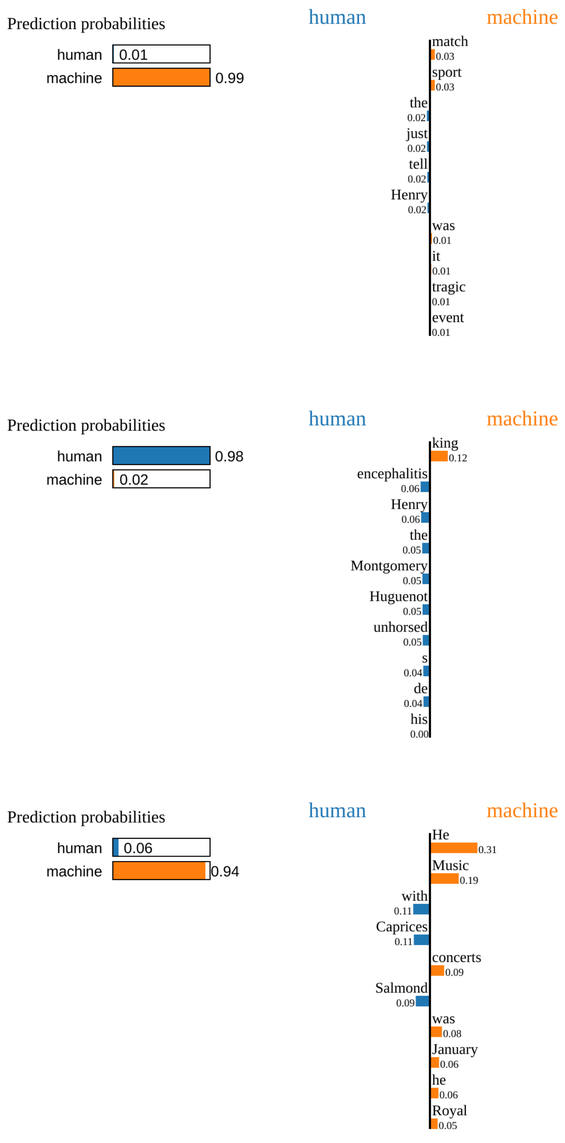
For each domain, we trained separate models and evaluated their performance both in-domain and cross-domain.

4. Results



Tiny-BERT (4.39M parameters) outperforms larger models with a 0.634 average accuracy across cross-domain setups, showing smaller models can generalize better for this task.

Reddit pretraining yields best cross-domain transfer for Tiny-BERT (0.723 accuracy), likely due to the diverse writing styles and topics represented in this conversational dataset.



True Positive

Text with highlighted words

Well, let me tell you, it was quite a tragic event for both Henry II and his opponent Gabriel de Montgomery. It all went down in 1559, during a jousting match at the Hotel des Tournelles in Paris. Henry was an experienced jouster, but Montgomery was a newcomer to the sport, and unfortunately, he was no match for the King's lance.

During one of their runs, Henry's lance struck Montgomery's helmet, shattering it and sending a jagged piece of wood into his eye and brain. The young man was rushed to a nearby hospital, but it was too late. He died just a few days later, leaving behind a wife and children.

As for Henry, he was devastated by the accident and reportedly went into a deep depression. He blamed himself for Montgomery's death and was haunted by guilt for the rest of his life. He even imposed a penance on himself, vowing to fast and do charitable works for the rest of his days.

The incident also led to changes in the sport of jousting. After Montgomery's death, many rules were put in place to make the

True Negative

Text with highlighted words

Henry died in a joust against the captain of his Scottish Guard, Gabriel, the Count of Montgomery. The fateful run occurred at the end of a tournament day, after Montgomery had almost unhorsed the king. Henry instead on another tilt. Montgomery's lance struck the king's helmet and shattered, with a long splinter running through the king's visor through his eye and into or near his brain. The king initially survived injury, and was attended to by two of the most celebrated physicians in Europe, Ambroise Paré and Andreas Vesalius. The queen, Catherine de Medici, ordered four prisoners executed with wood driven into their brains so that the physicians would have the chance to study the king's wound in detail on the corpses. Despite this rather extraordinary measure, the king deteriorated steadily. Vesalius' personal account is consistent with the development of meningitis or encephalitis. After 11 days, the king died.

During these 11 days, Montgomery is supposed to have come to the king's side, and asked to have his head and right hand cut off in punishment. The king told him that he had jousted well and bravely and that the accident was not his fault.

Following Henry's death, Catherine essentially ruled through a series of three of her sons. Montgomery retired to his estate in

False Positive

Text with highlighted words

William Edward Whitehouse (20 May 1859 – 12 January 1935) was an English cellist.

Career
He studied for one year with Alfredo Piatti, for whom he deputised (taking his place in concerts when called upon), and was his favourite pupil. He went on to teach at the Royal Academy of Music, Royal College of Music and King's College, Cambridge; his students included Felix Salmond and Beatrice Harrison, who both became closely associated with Edward Elgar. He played with violinist Joseph Joachim, and formed The London Trio with violinist Achille Simonetti and pianist Amina Goodwin. He edited Piatti's Caprices, with suggestions as to how his former teacher preferred them to be played.

External links
William Whitehouse
The Violoncello and the Romantic Era: 1820-1920: Part II — A Survey of Current Cello Teachers on Romantic Repertoire and Aesthetics

Our analysis shows human text contains more proper nouns and location references, while machine text exhibits generic vocabulary and simpler structures. These patterns reflect humans' personal knowledge versus models' statistical training, though they aren't consistent across all domains.

5. Conclusion

- **Model efficiency**: Tiny-BERT (4.39M params) achieves best cross-domain accuracy (0.634).
- **Domain transfer**: Reddit pretraining yields optimal generalization (0.723).
- **Linguistic markers**: Human texts use proper nouns; machine texts favor generic terms.
- **Parameter scaling**: More parameters don't improve detection performance in cross-domain settings.