CSC490 A1

Benson Li	1007815376
Yuchen Wang	1010441595
Xuanyi Lyu	1009343266
Zheyu Zheng	1008979580

September 2025

1 Interest Statements

With the rapid advancement of AI, public discourse is increasingly influenced by low-quality AI-generated content. This can lead to issues such as misinformation, spam, and even fraud, which threaten the credibility of online information and undermine the trust users place in digital communities. To address this challenge, our team is developing a machine learning model that can detect AI-generated text on social media platforms, helping users distinguish genuine human expression from automatically generated content. Our team members are interested in contributing to the following areas.

- **Benson Li:** Designing and training transformer models tailored to Twitter's unique short-text format, while optimizing model performance, ensuring robustness against noisy inputs, and enabling efficient real-time analysis for end users.
- **Yuchen Wang:** Collecting and pre-processing large-scale text datasets across diverse contexts to ensure training on high-quality, representative data with reduced noise and bias, leading to more robust performance on unseen inputs.
- **Xuanyi Lyu:** Evaluating the detection model against real-world adversarial attacks, such as watermarking and paraphrasing, to assess and improve robustness, while also exploring how detection results can be presented in a clear and user-friendly manner to enhance transparency and trust.
- **Zheyu Zheng:** Establishing baseline models (RNN, LSTM) for benchmarking against transformers, as while as conducting ablation studies on preprocessing choices and help with experimental design and documentation of results to ensure consistent and reproducible evaluation across all models.

2 Landscape Analysis

Item	**Description**	**Commentary**
Turnitin	A similarity-checking platform widely used in academia, now integrating AI detection with transformer-based deep learning.	Key gaps: Primarily focused on academic content, and may struggle with more diverse content such as social media.
Originality.ai	A detection tool designed for publishers, website owners, and content creators, mainly for blogging, SEO, and online publishing.	Key gaps: Limited to web publishing and SEO contexts; effectiveness in social media and multilingual text remains underexplored. Studies show that paraphrasing or "humanizing" AI text reduces detection accuracy. [4, 5, 8]

Item	Description	Commentary
GPTZero	An AI detection model that analyzes text using features such as perplexity and burstiness.	Key gaps: Accuracy varies significantly with text length; shorter or noisier passages are more likely to be misclassified compared to longer essays. [6, 7]
Winston AI	A high-accuracy AI detector (claiming 99.98%) that can detect paraphrased and "humanized" content, offering an AI Prediction Map to visualize sentences.	Key gaps: Focus on document-level analysis and high-stakes accuracy makes it less suitable for high-volume, probabilistic needs like social media.
Pangram	An AI detection tool emphasizing near-zero false positives and fairness for non-native speakers, trained on diverse text across domains.	Key strengths: Outperforms other detectors across news, email, and reviews, making it a strong baseline for comparison with other models.
RoBERTa	A robustly optimized BERT pretraining approach that removes the NSP objective, applies dynamic masking, and trains longer on more diverse data.	Key strengths: Widely adopted as a backbone encoder; its contextual representations improve classifier accuracy in distinguishing human- vs. AI-generated text.
GLTR	(Giant Language Model Test Room) uses statistical methods to visualize word predictability, showing how AI text tends to use more common words than human writing.	Key strengths: Provides an interactive visualization that helps humans intuitively identify overly regular patterns in machine-generated text.
DetectGPT	A zero-shot detection method that leverages probability curvature, based on the hypothesis that AI-generated text lies in regions of negative curvature in log-probability.	Key strengths: Requires no training data or classifier; however, struggles with informal or mixed-style text, making it less reliable for short, fragmented social media posts. [3]
SynthID	A watermarking system by Google DeepMind that embeds imperceptible digital watermarks into AI-generated text, images, audio, and video.	Key strengths: Ensures traceability of AI content by embedding invisible watermarks that remain detectable even after modification. [1]
Copyleaks	An enterprise-level AI detection platform supporting multiple languages and integration into workflows, widely used in education and industry.	Key strengths: Enables large-scale detection across document collections, suitable for organizations requiring automated, enterprise-grade verification.
DeTeCtive	A neural-network-based AI content detector that learns from large datasets to capture subtle differences in style, coherence, and word usage.	Key strengths: Achieves state-of-the-art performance across benchmarks; in zero-shot cross-domain tests, it outperforms SimCSE-RoBERTa. Introduces TFIA mechanism, improving out-of-distribution detection with minimal adaptation data. [2]

3 Project Outline

3.1 Problem Statement

With the proliferation of AI-generated content on social media platforms, particularly Twitter, users are increasingly exposed to misinformation, spam, and deceptive automated accounts. Large Language Models (LLMs) such as ChatGPT and Bard are now widely used to generate text at scale, making it difficult to distinguish between human and AI-generated content. This poses a direct threat to the credibility of online information and undermines public trust in digital communication platforms.

3.2 Proposed Solution

Our team proposes to develop a transformer-based detection system specifically designed for Twitter's short-text format. This system will classify tweets as either human-written or AI-generated, while also assessing robustness against adversarial techniques such as watermarking, paraphrasing, and other obfuscation methods. By combining large-scale, high-quality datasets with state-of-the-art transformer models like BERT, our tool aims to provide accurate, transparent, and actionable information to end-users.

3.3 High-Level Technical Approach and Milestones

Technical Approach

- **Dataset Collection & Preprocessing:** Combine large-scale datasets of AI-generated tweets (e.g., GenAITweets10k) with curated human-written tweets (e.g., Sentiment140) to ensure balanced and representative samples. Clean and tokenize data with specialized tools (e.g., NLTK TweetTokenizer and BERT tokenizer) to handle emojis, hashtags, and Twitter handles.
- **Model Development:** Train baseline models (RNN and LSTM) for benchmarking, then fine-tune BERT-based transformer models with both frozen and flexible weight configurations to explore transfer learning performance.
- **Adversarial Robustness Testing:** Evaluate model resilience against real-world attacks such as watermarking, paraphrasing, and prompt-injection to ensure the system maintains high accuracy under adversarial conditions.
- **User-Focused Output:** Design an interface or reporting format that clearly communicates detection results (confidence scores, explanations) in a user-friendly way to promote transparency and trust.

Milestones

- Milestone 1 (Weeks 3–4): Collect and preprocess AI-generated and human-written tweets; ensure dataset balance.
- Milestone 2 (Weeks 5–6): Train baseline RNN and LSTM models; evaluate and document performance.
- Milestone 3 (Weeks 7–8): Fine-tune BERT models (frozen and unfrozen weights) and benchmark key performance metrics, including accuracy, precision, and recall.
- Milestone 4 (Weeks 9–10): Conduct adversarial robustness tests (watermarking, paraphrasing, prompt-injection); refine detection thresholds.
- Milestone 5 (Weeks 11–12): Develop and test user-facing interface for clear display of detection results; finalize prototype, evaluation, write-up, and presentation.

3.4 List of Unknowns to Investigate

- How well will the model generalize to new or unseen AI models that generate text differently from the training data?
- What detection threshold best balances false positives and false negatives in a real-world setting?
- How to design an interface that conveys model uncertainty and explanations without overwhelming the user?
- What ethical and privacy implications arise when collecting and analyzing large volumes of social media data?
- Can detection remain robust when faced with deliberate obfuscation techniques such as adversarial paraphrasing or watermark removal?

4 Project Press Release

New AI Detection Tool "TweetVerify" Helps Twitter Users Distinguish Human vs AI Content — 98% Accuracy

A team of machine learning researchers today announced **TweetVerify**, a cutting-edge AI detection system built to identify artificially generated content on Twitter. The tool addresses the growing challenge of separating authentic human expression from AI-generated text that threatens the credibility of online discourse.

As AI models become increasingly sophisticated, social media platforms face new risks from misinformation, spam, and automated accounts. **TweetVerify** represents a breakthrough, achieving 98% accuracy on Twitter's short-text format while remaining resilient to adversarial attacks. The system is lightweight, scalable, and designed to handle millions of tweets daily, making it practical for deployment at platform scale.

"Most existing AI detectors were not designed for the realities of social networks," said Benson Li, the lead researcher. "Twitter's mix of short posts, emojis, hashtags, and informal language creates unique challenges — TweetVerify was built from the ground up for this environment."

Key Features and Benefits

- **Social Media Awareness:** Purpose-built for Twitter, excelling at analyzing short tweets, threads, hashtags, and mentions where generic detectors often fail.
- **Resilience:** Maintains high accuracy even against paraphrasing, watermark removal, and other adversarial tactics.
- **Transparency:** Provides clear confidence scores and human-readable explanations, helping users understand and verify every decision.
- **Efficiency:** Delivers instant results without slowing platforms or disrupting user engagement, ensuring a seamless experience.

Impact

TweetVerify combines compact transformer models, Twitter-aware tokenization, and adversarial testing to deliver a detection tool designed for real-world use. Unlike many academic benchmarks, TweetVerify prioritizes deployability, interpretability, and trust.

Early pilots show strong impact across industries. "As a reporter, speed and accuracy matter — TweetVerify is now an essential part of our verification workflow," said XXX, investigative journalist. Content creators report stronger audience trust after adopting the tool, while educators are using it as a hands-on resource for teaching digital and media literacy. Platform stakeholders note the potential for TweetVerify to strengthen public trust in social media ecosystems by curbing the spread of synthetic misinformation.

Appendix — Press Release Iterations

The press release went through several iterations to balance technical accuracy with accessibility:

- **Iteration 1:** The first draft was too technical and filled with jargon (e.g., discussions of loss functions, fine-tuning steps, and hyperparameter choices). It lacked a clear headline and was inaccessible to non-technical readers.
- **Iteration 2:** This version addressed the accessibility issue by simplifying technical language and adding a clear, attention-grabbing headline. However, it narrowed the focus too much to journalists and fact-checking, missing broader benefits for educators, creators, or everyday Twitter users.
- **Final Version:** Built on iteration 2 by expanding the audience, adding relatable fictional quotes, and highlighting measurable accuracy (98%). It framed TweetVerify as an accessible, trustworthy tool that empowers diverse user groups while maintaining a solution-focused narrative.

References

[1] See A. Ghaisas S. et al. Dathathri, S. Scalable watermarking for identifying large language model outputs. *Nature*, Oct.2024.

[2] Xun Guo, Shan Zhang, Yongxin He, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. Detective: Detecting ai-generated text via multi-level contrastive learning. 2024.

[3] Praveen Rao Hulayyil Alshammari. Evaluating the performance of ai text detectors, few-shot and zero-shot classifiers on human-ai generated content. *arXiv preprint arXiv:2507.17944*, 2025.

[4] Fadi Sibai Karim Hesham Shaker Ibrahim, Dhari Abdullah Alotaibi. The robustness of ai-classifiers in the face of ai-assisted plagiarism. *ResearchGate Preprint*, 2025.

[5] Keshav Krishna et al. Dipper: Paraphrasing to evade ai-text detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[6] Michael Perkins et al. Evaluating the efficacy of ai content detection tools: An analysis of openai's classifier and gptzero. *International Journal for Educational Integrity*, 19(3):1–15, 2023.

[7] Mehmet Dikn Selin Dik, Osman Erdem. Assessing gptzero's accuracy in identifying ai vs. human-written essays. *arXiv preprint arXiv:2506.23517*, 2025.

[8] Mehrdad Saberi Shoumik Saha Soheil Feizi Yize Cheng, Vinu Sankar Sadasivan. Adversarial paraphrasing: A universal attack on ai text detectors. *arXiv preprint arXiv:2506.07001*, 2025.