Detection, Validation and Analysis of Deepfaking in Social Media

Cole Ouellette[1 [0000−0002−1835−5892]], Brian Almaguer[1 [0009−0007−5352−6961]], Xiaoyan Sun[1 [0000−0002−0321−2338]], and Jun Dai[1 [0000−0002−6890−6429]]

Worcester Polytechnic Institute, Worcester MA 01609, USA
{ceouellette,balmaguer,xsun7,jdai}@wpi.edu

Abstract. As AI continues to gain momentum and capture media attention, the prevalence of AI-generated voice, video, and image deepfakes on social media is rising dramatically. On some platforms, the AI attribution rate is estimated to be as high as 38.95%, and continuing to increase month over month. Despite significant efforts in the literature to detect and mitigate these synthetic media, a substantial amount of deepfaked content still circulates undetected. This paper investigates methods for identifying, classifying, validating, and analyzing deepfakes in the digital landscape. Furthermore, it investigates the potential of leveraging contextual insights through the analysis of user interactions with AI-generated content, and proposes a new technique to enhance the detection and mitigation of deepfakes across social media platforms. By outsourcing part of the overall detection computation to a distributed base of users, interaction-based detection could prove to be a successful tool in reducing the harmful effects of AI-generated deepfakes online.

Keywords: Deepfake detection · Social media · Interaction-based detection · Generative AI · Logistic regression.

1 Introduction

As generative AI tools continue to spread into the consumer market, creating AI-generated media has never been more accessible. In the years since OpenAI's *ChatGPT* [27] first hit the market, work adoption of generative AI has been as fast as the personal computer, and overall adoption has been faster than either PCs or the internet [9]. With the availability of large language models such as *ChatGPT* and *DALL-E 3* [28], a serious concern is raised about such tools being used to create fictitious, synthetic content. The potential for deepfakes to circulate misinformation, fake news, propaganda, and distrust over the internet is at a level never before seen [33].

Though methods exist to identify AI-generated content by examining the potentially deepfaked content itself, the online era presents the unique ability to leverage potentials for a new detection technique, enabled by the mass amount of users interacting with content on social media. Common flaws in AI-generated content including incoherence, insufficient evidence-based claims, and vocabulary diversity in text-based content [24], as well as irregularities in eyes, visible

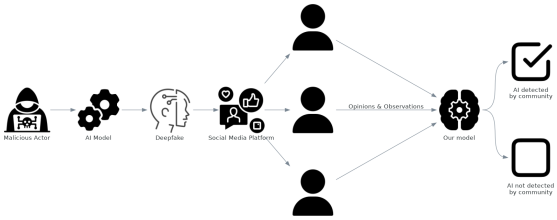


Fig. 1: The distillation process of a deepfake on social media, ultimately picked up and identified by this project's methods.

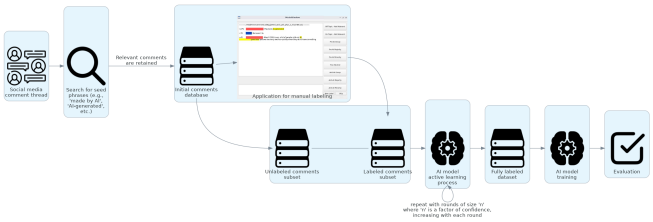


Fig. 2: Project detection pipeline. This work presents an AI model capable of detecting synthetic posts on social media by analyzing subsequent user interactions. We build an initial database of submissions and comments by searching comments for seed texts indicating doubt in the content's legitimacy. A subset of the database is manually-labeled by researchers to indicate if the commenter believes the parent content is human- or AI-generated. The labeled data is then used in active training an AI model to label data, repeating with rounds of increasing size with each round.

distortions, and non-uniformity in image-based content [25] can be perceived by humans and used to attribute content to an AI-generated origin. By outsourcing part of the overall detection computation to a distributed base of users, interaction-based detection could prove to be a successful tool in reducing the harmful effects of AI-generated deepfakes online.

Figure 1 shows the high-level timeline of the deepfake distillation process from creation to detection. This paper proposes a methodology to intake a series of user interactions and comments underneath a given social media posting and output an attribution indicating that the post is either human- or AI-generated. More specifically, we build an initial database of social media submissions and comments by automatically searching comments for key subtexts containing English phrases hypothesized to be used in expressing doubt of the content's authenticity. A subset of the initial comments database is manually-labeled as to whether the commenter believes the parent content is legitimate. The labeled data is then used in active training an AI model to assign labels to comments, repeating with rounds of size 'n' where 'n' is a factor of confidence, increasing with each round until confidence exceeds 99%. This process can be seen in Figure 2.

Once a fully labeled dataset is realized, the model is evaluated on its accuracy in identifying comments indicating AI-generated content and the hypothesis is evaluated on its ability to detect AI-generated content from an interaction-based approach.

This paper's contributions are as follows:

- Developed a PyQt5 GUI application to streamline comment thread creation and visualization for manual labeling. Developed for Reddit thread structure, the application can be easily modified to support data from other social media applications.
- Constructed a first-of-its-kind dataset of manually-labeled contextual social media submission data providing information on whether the commenters involved suspect AI origins.
- Designed a data processing pipeline and logistic regression model to predict probabilities that a social media submission was deepfaked, using human-generated community insights. All pipeline components are built to be platform-agnostic and can be utilized to make predictions for content on any social media platform.
- Built an alternative post-engine processing pipeline that builds upon the original model, enabled by large language models to make decisions about the content's relevance and likelihood of being deepfaked, thus weeding out false positive results and reducing unnecessary computation time.
- Evaluated base detection model and LLM-enabled model performances. The base interaction-based approach yielded a 56.99% accuracy with a 58.33% precision. The LLM-enabled model slightly outperformed the base model with 58.82% accuracy and 67.16% precision.
- Propose methods for future improvements, including a theme-based detection strategy utilizing a novel database of "known-falsehoods" approach.
- Independently confirmed data regarding humans' inability to consistently detect AI, paralleling results discovered in recent studies. While the novel methodology that this paper introduces did not prove to be a reliable means of detecting AI-generated content online, the replication of human benchmark results in a previously unexplored social media context reinforces the need for an accurate, hands-off method to detect synthetic content online.

2 Background

This section explores leading methods for identifying AI-generated content across several mediums and a ground-truth methodology for distinguishing bot accounts from human accounts online.

2.1 Current Deepfake and AI-Generated Content Detection Methodologies

Leading methods for identifying AI-generated content examine the content itself without surrounding contextual information to influence the decision. By target-

ing the content itself, methodologies are generally restricted in their application to a specific media type.

Detecting AI-Generated Text. Document-level AI-generated text detection methods fall short in cases where users coauthor their papers with AI, creating a work comprised of sentences written by humans and modified by AI or vice versa [34]. Wang et al. propose a new strategy to detect AI-generated text on a fine-grained sentence-level, better suited to attribute a subsection of a larger text to AI. The researchers' methodology segments an input text and assigns human/AI attribution to each token, selecting the most frequent label in the sequence for the overall sentence label.

Detecting AI-Generated Images. Recent advances in AI image-generation technologies result in new models capable of creating photorealistic images regularly hitting the market. Baraheem and Nguyen [6] utilize a trained convolutional neural network to separate Generative Adversarial Network (GAN)-generated images from real human-generated images. Baraheem and Nguyen's model correctly attributes images as human or AI, with an average accuracy of 92.3% across three different datasets. By training the model to identify common flaws, distortions, and artifacts left behind from the GAN image generation process including pixel anomalies, artificial fingerprints, and spectral artifacts, this methodology can be generalized to any image generation model so long as it produces images with similar flaws. Epstein et al. [18] also present a platform-agnostic way to detect AI-generated images with a convolutional neural network incrementally trained on 14 popular generative models released between 2020 and 2023 to simulate a real-world learning setting where the detection methods can be applied to new models that have yet to be released.

Cozzolino et al. [15] propose a technique based on features extracted from the image encoder of a pre-trained vision language model, the Contrastive Language-Image Pre-Training (CLIP). By feeding CLIP with both real and synthetic image pairs, Cozzolino et al. can extract the feature vector outputs to then design a linear SVM classifier around.

Detecting AI-Generated Videos. Yu et al. [36] propose an image-segmentation approach for detecting deepfake facial videos. The researchers devised a methodology to segment a frame of an image into patches, isolating facial features to closely study each segment for artifacts of deepfaking.

Mitra et al. [26] propose a deep neural method for detecting social media deepfake videos with a 92.33% accuracy against FaceForensics++ [29] and Deep-Fake Detection Challenge [17] datasets. To reduce the computational impact of detecting videos at a social media scale, Mitra et al. apply a key video frame extraction technique on each video and take those key frames as an input for processing.

The Push for a Social Media-Centric Approach. In contrast to the existing aforementioned methods to detect AI-generated content online, the work this paper presents utilizes contextual user insights found on social media platforms to enable detection on a large scale, irrespective of content medium. Unlike previous efforts by Wang et al., Baraheem and Nguyen, and Epstein et al., this project's methodology does not rely on any particular-model binary detection requiring the model to train on a specific model's identifiable artifacts. Instead, the interaction-based approach leverages user insights meaning that, as human-visible artifacts are identified in the outputs from newer models, the community is able to dynamically adjust and pick up on the clues attributing a post to AI. In a similar nature to the work done by Mitra et al., extracting only key video frames for processing rather than every frame in a video to minimize required computations, the interaction-based methodology that this paper presents only processes social media submissions indicated by commenters to perhaps be AI-generated. This outsourcing of processing power to the broad, distributed judgment of a social media user base eliminates the need to process each submission online.

2.2 Ground-Truth Approach for Distinguishing Humans from Machines

In an effort to detect GitHub bot accounts masquerading as human users, Golzadeh et al. [20] created a ground truth dataset capturing 5,000 GitHub accounts with authenticity labels classifying each account as human or bot. This ground-truth dataset enables researchers to later train an automated classification model for detecting bots based on their commenting activity in GitHub issues and pull requests.

To generate the dataset, Golzadeh et al. gathered a random sample of commenters, a subset of commenters previously believed to be bots, and another subset of GitHub accounts with usernames matching a curated list of substrings believed to be commonly utilized by bot accounts. Researchers then manually and iteratively reviewed each account's GitHub commenting activity and rated them on a multifactor scale identifying an account as either a bot or human, with an associated confidence value with each decision.

2.3 Related Fields of Study

An interaction-based method for AI detection has roots in other areas of study, namely social media identity deception [16,10,4], bot detection in social media [3,31,21,23], and sentiment analysis [8,19,7,12,5,13]. This research provides the backbone for the interaction-based methodologies that this paper presents.

3 Implementation

This section discusses the design goals necessary in developing a thorough interaction-based detection model and the implementation details of doing so. It also high-

lights how the model is evaluated, its performance results, and some designs to further leverage identified contextual insights and improve future detection performance.

3.1 Design Goals

We designed the interaction-based detection model around the following criteria, in no particular order. By accomplishing these design goals, a detection model's efficiency and performance can be maximized, and the methodology presented can be applied to any number of social media contexts.

- **Goal 1:** ***Platform Agnosticism.*** An interaction-based detection model that examines different social media contexts needs to be designed to work irrespective of the social media platform being examined. A successful strategy that works to detect AI-generated content on one platform, such as Reddit, should, with little modification, be made to work well on another social media platform where similar user insights may be drawn from.
- **Goal 2:** ***Intent-Aware Detection.*** In certain online social media communities, conversations about various AI topics and AI-generated content showcases are completely regular, and welcomed, occurrences. To eliminate false positives, a detection algorithm must be able to differentiate between benign conversations discussing AI, submissions in which the author openly acknowledges using AI, and real instances where a bad actor is trying to pass off synthetic content as human-generated.
- **Goal 3:** ***Consistent Performance Over Variations in User Engagement.*** Social media user engagement levels vary wildly between submissions, but it is important that an interaction-based detection model performs just as well in submissions with a handful of user interactions as it does in those with thousands.
- **Goal 4:** ***Mixed-Model Binary Detection Formulation.*** Wang et al. claim that when creating a detection model, there are generally three different formulations to consider: particular-model binary detection, mixed-model binary detection, and mixed-model multiclass detection [34]. Models classified under a particular-model binary detection formulation are trained to detect whether a piece of content was produced by a *specific* known AI model. Instead of looking for origins in a particular AI model, mixed-model binary detection models predict whether content was produced by *any* AI model. Going even further, mixed-model multiclass detection formulations detect AI model origin and attempt to attribute it to a specific model, tracking down the exact culprit. For the purposes of this paper, we care only about attributing content to human- or AI-generated origins, so a mixed-model binary detection formulation is the minimum level of specificity required.

3.2 Design Overview

This project explores an interaction-based detection strategy to identify AI-generated deepfake content posted on social media platforms. We collect sub-

mission and comment data from social media platforms to then label and train a model to identify comments that are raising suspicion around a submission's authenticity.

Data Collection. In order to develop our deepfake detection model, we must collect and analyze existing user responses to AI-generated content in social media communities. For our methodology to apply regardless of the social media platform, our data needs to utilize platform-agnostic metrics.

In its most simplistic form, a submission can be boiled down to a piece of content, with an author, a content score, a list of comment threads in reply to the submission, as well as a title, a description, or both. Similarly, comments have authors, a content score, a parent submission, a list of subsequent child comment threads, and a body text.

While many social media platforms match this basic structure, this work chose to analyze user engagement on Reddit due to their well-documented Python API wrapper, *PRAW* [2], and the years' worth of data dumps made readily available through the *Pushshift* API [32], recording every submission and comment made to the most popular 40,000 Subreddits each month. Though the project is designed on the back of Reddit's data structure, the methodology presented can easily be adapted to fit the unique structure conventions of other social media platforms, such as TikTok or X.

Initial data was gathered from the *Pushshift* dumps, collecting all comments and submissions from the top 40,000 Reddit communities month-to-month in the calendar year 2024 to ensure sufficient data availability. For processing comment threads, this archival data was supplemented with *PRAW* queries to capture all more recent user engagement and fill in any information gaps, where needed. The archival data amounted to 1.2 TB in data storage for 2024 alone. To make data querying more efficient and streamline data manipulation, comment and submission data was uncompressed from their original file format and stored in an SQLite database, using unique comment and submission IDs as primary keys for fast retrieval. The SQLite database schema is as follows:

```
CREATE TABLE comments (
    name TEXT PRIMARY KEY,
    link_id TEXT NOT NULL,
    parent_id TEXT NOT NULL,
    author TEXT NOT NULL,
    body TEXT NOT NULL,
    permalink TEXT NOT NULL,
    score INTEGER NOT NULL,
    created_utc DATE NOT NULL
);
```

Enabled by our hypothesis that a subset of people will identify and call out AI-generated content online, this information was then used to explore user interactions with submissions as a metric for detecting deepfakes, building feature

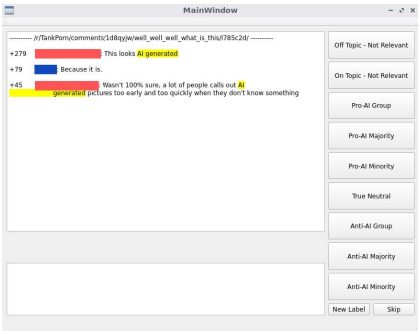


Fig. 3: Python application created to interface with the SQLite database and manually assign labels to comment threads.

vectors that hold key structural details about the post and the surrounding contextual user-engagement details. The feature vectors and their contents are discussed in more detail in Section 3.2.

Filtering on the comment body identifies a subset of comments in which the user has claimed the parent submission is AI-generated. We then use subsequent replies and the reply's user-voted score as a metric to gauge if the comment is likely correct in its assessment, and modify the submission's feature vector.

Data Annotation. In order to train a model to detect AI-generated content, a custom Python application built on top of PyQt5 [35] is used to interface with the SQLite database, retrieve the comment threads, and manually assign a label accordingly, as seen in Figure 3. One by one, the application retrieves comment threads from the database and highlights the seed text that initially indicated the thread's potential relevance. By reviewing the thread's comments, the parent post's subject matter, and the user-voted score of each comment, researchers assign a label indicating the thread's AI attribution, as claimed by the commenters involved.

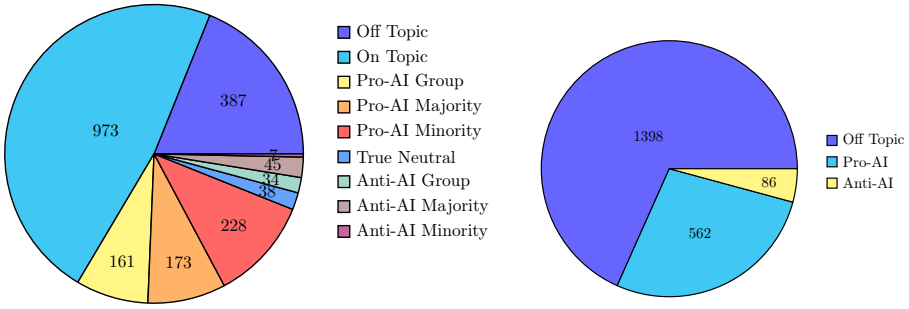
When developing labeling schema for the training process, the research team elected to create a Likert-type scale indicating the thread's relevance that could be adapted to fit our needs in any condition later on. This resulted in two distinct, fine- and coarse-grained label distributions, the breakdowns of which can be seen in Figure 4.

1. **Fine-Grained Label Distribution.**
   - *Off Topic – Not Relevant*: The seed texts used to select threads for labeling include phrases such as "by AI." This may sometimes result in falsely highlighting comments that contain the seed text as an irrelevant substring, "*by ai*rplane" for example. This label is used to collect such threads deemed irrelevant to the detection model.
   - *On Topic – Not Relevant*: With the increased adoption of AI in everyday use, it is not uncommon to discover discussions about AI taking place

on social media. This category is used to capture conversations in which the commenters involved are discussing AI generated content in some manner, but without any questioning of the parent contents' legitimacy.
- *Pro- & Anti-AI Group/Majority/Minority*: This categorization range differentiates between the decision and confidence strengths in each conversation. If the commenters generally agree that the submission is created by an AI model, it is filed under the Pro-AI umbrella. Likewise, if the observed interactions signify that the general consensus is that the submission is human-generated, it falls under Anti-AI. Depending on the engagement's conviction strength, the thread can be sub-categorized under Group, Majority, or Minority. A group label means that every commenter involved in a given interaction is in agreement with the submission's attribution and each opinion is generally well-received, as judged by the comments' user score. If there are differing opinions in the thread, the decision between one category's majority label and the other's minority is left up to researcher judgment, guided by which party made the strongest claim.
- *True Neutral*: In the case that both arguments are present with equal conviction, the neutral label is reserved for threads where looking at the user interaction data presents no clear decision.

2. **Coarse-Grained Label Distribution.** Unlike with fine-grained label distribution, the coarse-grained approach was designed to sort annotated comment threads into data more fit for a binary classification model. Threads labeled and found to fall into any of the Pro-AI Group/Majority/Minority collections were combined into one Pro-AI label, those assigned Anti-AI labels were combined into a general Anti-AI, and all else was deemed Off Topic. Anti-AI content was kept separate to enable future classification choices, but the first classification model in this project's interaction-based detection pipeline treated Pro- and Anti- content alike to distinguish between AI accusations and everything else, leaving the ultimate AI- or human-generated classification decision to our feature-vector-powered statistic model.

For fastest results on a large scale, this work uses manually-labeled data and an active training approach to develop a model capable of replicating this comment labeling procedure and predicting the community's classification decision from discussions within a comment thread. In training the model, we gradually increased the pool of manually-labeled conversations and evaluated the model's performance in making predictions by observing the model's results. As the training set surpassed 1,400 comments, the model demonstrated a large increase in prediction accuracy. Manual labeling continued until the labeled set reached a size of 2,044 annotated comment threads to ensure statistical significance in the model's ability to distinguish Pro- and Anti-AI community consensuses. With a labeled dataset containing 2,000 manually-labeled comment threads, an active training pipeline taught the model to predict thread classifications with increased confidence. Each round of active training iterated through a size 'n' subset of the remaining seed-selected comment threads, where 'n' is a factor of



(a) Fine-grained labels. (b) Coarse-grained labels.

Fig. 4: Label distribution of 2,044 manually-annotated comment threads.

confidence, increasing with each round until the confidence exceeds 99%. The process of generating an initial comments dataset from seed phrases, manually labeling, and training the classifier with active learning can be seen in Figure 2.

Generating Feature Vectors. For the model presented in this paper to properly examine a submission's contextual insights and attribute it, we must first generate a feature vector to record and analyze. Feature vectors' contents are flexible, and the information they contain is passed to the statistic model for the final step in the detection pipeline. Our feature vectors combine binary and cumulative metrics observed in a post's subsequent interaction signals, chosen by the research team as plausible predictor factors of a submission's authenticity:

- *Art Community?*: A binary value indicating whether the post was submitted in a community centered around artistic themes. This hypothetical predictor was birthed out of the hypothesis that synthetic content could be more prominently shared in art communities due to the widespread availability of generative AI as a creative tool.
- *Sensitive Topic?*: Like the *Art Community?* predictor, this variable is a binary score recording whether the target submission is classified as a sensitive topic. If a malicious actor was attempting to spread misinformation and distrust online, a likely way to do so is by creating synthetic content, pushing a narrative that evokes strong emotions, reactions, and discourse. By observing a post's title, body, and community, we can determine if the submission's topic should be deemed sensitive.
- *Active but Downvoted?*: In the case of users identifying a more apparent deepfake claiming to be AI generated, one reactionary behavior that we observe is a large amount of subsequent discussion beneath a post 'disliked' by the community. This behavior does not solely exist in response to deepfakes, but we believe it could be a possible predictor.
- *AI Confirmed by Author?*: Though we consider transparent AI usage outside the scope of this interaction-based detector, this binary predictor factor

accounts for the instances where a post may initially be passed off as human-generated, but the author later concedes to using AI in its creation.
- *AI Comment with Positive Score & AI Comment with Negative Score*: Each time that a comment notes the content being human- or AI-generated, one of these predictors is incremented. These variables are used to account for the silent consensus shared by the non-vocal users engaging with a post. If a comment is well received and has a positive user-voted score, the value of *AI Comment with Positive Score* will increase. Otherwise, and the comment is met with disagreement through a negative user-voted score, the value of *AI Comment with Negative Score* will increase.
- *Top Level Comments Indicating AI*: Each time a top level comment is found that expresses doubt in the parent being human-generated, this value increases. If many users are individually attributing the content to AI origins, it should be a sign that influences the model's prediction.
- *Child Comments also Discussing AI*: Building off of the insights from the *Top Level Comments Indicating AI* predictor, this predictor increments whenever a child comment is found to agree with a parent that the submission is AI generated. Whereas finding an accusatory top level comment is relatively straightforward using keyword searches, identifying child comments in agreement requires sentiment analysis to detect.

The algorithm this work follows to take a social media post and construct a vector containing these predictor factors is shown in more detail in Figure 5.

Statistical Prediction Model. Finally, a statistical prediction model is also trained on the manually-labeled coarse-grained comment thread data to process a feature vector and make a prediction whether a submission is human- or AI-generated. Though many options for prediction models exist, this work chose to implement a logistic regression model. Logistic regression is a great fit for this use case due to its strength in predicting a dependent variable from independent variables that may be associated with the outcome. When fed a set of independent variables, the model outputs a prediction score signifying the likelihood that the dependent variable would be positive, given the independent variables.

For a logistic regression model to be deemed viable, the following three criteria must be satisfied:

1. The model must yield accuracy greater than 60%.
2. P values of all coefficients (except the constant) must be less than 0.1, indicating 90% significance.
3. P value of the overall model must be less than 0.1, indicating 90% significance.

One common point of failure for logistic regression models comes from having an imbalanced training set. When one class is largely overrepresented, the optimizer is biased and has a tendency to skew predictions toward the majority class. The manually-annotated data, discussed in more detail in Section 3.2, consists of 562 positive entries and 1,484 negative entries. Due to the majority class

Fig. 5: Feature vectors generation algorithm for each submission.

```
let featureVector ← {
    "Art Community?": 0,
    "Sensitive Topic?": 0,
    "Active but Downvoted?": 0,
    "AI Confirmed by Author?": 0,
    "Top Level Comment Indicates AI": 0,
    "AI Comment with Positive Score": 0,
    "AI Comment with Negative Score": 0,
    "Child Comments also Discussing AI": 0}
if Parent Post Does Not Claim AI then
    if Post Belongs to Art Community then
        featureVector['Art Community?'] ← 1
    end if
    if Post is a Sensitive Topic then
        featureVector['Sensitive Topic?'] ← 1
    end if
    if Post Score is 0 and Post Comment Count > 1 then
        featureVector['Active but Downvoted?'] ← 1
    end if
    for Each Top-Level Comment do
        if Comment Claims AI then
            featureVector['Top Level Comment Indicates AI'] ← featureVector['Top Level Comment Indicates AI'] + 1
            if Post Author Claims AI then
                featureVector['AI Confirmed by Author?'] ← 1
                break
            end if
            if Comment is Upvoted then
                featureVector['AI Comment with Positive Score?'] ← featureVector['AI Comment with Positive Score?'] + 1
            else
                featureVector['AI Comment with Negative Score?'] ← featureVector['AI Comment with Negative Score?'] + 1
            end if
            for Each Child Comment do
                if Child Claims AI then
                    featureVector['Child Comments also Discussing AI'] ← featureVector['Child Comments also Discussing AI'] + 1
                end if
            end for
        end if
    end for
end if
```

being nearly three times the size of the minority class, the logistic regression model is affected by this bias. To mitigate the impacts of the skewed data, we downsample the training set by randomly selecting 562 items from the negative class to retain and train with and reduce optimizer bias. Once the dataset is balanced, a 60:40 train/test split divides the dataset up for model training and evaluation.

3.3 Performance Evaluation

To understand how this novel interaction-based detection methodology performs, the model is run against a subset of our labeled data to compare its predictions with the expected values. This performance is used to determine the model's initial efficacy, and its shortcomings are examined to determine potential enhancements in the detection pipeline.

Evaluation Methodology. In binary classification problems such as the one this work tackles, accuracy, precision, recall, and F1 score each emphasize different aspects of a model's performance and can be used to evaluate a model. Accuracy provides a good rough idea of how well a model generally performs, while precision and recall are better metrics in situations where false positives or false negatives are more costly, respectively, and F1 score provides a better overall assessment of model performance when the data distribution is skewed.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

While a model's accuracy is simple to calculate, summing up the number of correctly predicted values and dividing by the total number of predictions, it might not always be the most suitable metric for judging a model's overall performance, especially when the data is imbalanced [22]. When the negative class is dominant, as it is in this case where a social media post is far more likely human-generated than by AI, simply predicting negative the majority of the time yields a high accuracy. Instead, we must analyze the threat model and identify what the model's performance needs are to ensure they are prioritized.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

In situations where false positives are costly, a high precision score means that we can be sure about positive predictions. In other words, when a model achieves high precision, there is a strong confidence that the positive predictions really are positive. However, false negatives are completely ignored, so there is no guarantee that positives are not going undetected. Predicting negative in the vast majority of cases is likely to yield a high precision value without being a useful judge of the model's performance.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

When evaluating a model by its recall, a high recall value means that you can be confident that all actual positives present in the sample were correctly identified. Recall measures the proportion of true positives compared to the total amount of actual positives. This metric can be important in cases where it is critical that all positives are identified, but is less useful when false positives are something we care about.

$$F1\,Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Finally, the F1 score. Unlike accuracy measuring the overall correctness of predictions with the arithmetic mean, the harmonic mean of precision and recall used to calculate the F1 score is better to use when correctly identifying positives is important, but no type of error is inconsequential. Using an F1 score to evaluate a model's performance is helpful in cases where false positives and false negatives are equally important, and the dataset is imbalanced.

Table 1: Logistic regression results of the interaction-based detection pipeline. The model demonstrates that the "AI Comment with Negative Score," "AI Comment with Positive Score," "Child Comment also Discussing AI," and "Top Level Comment Indicates AI?" independent variables have an impact on the outcome with greater than 90% significance.

Dep. Variable	Researcher-Assigned Label	**No. Observations**	408			
Model:	Logit	**Df Residuals:**	404			
Method:	MLE	**Df Model:**	3			
Time:	20:31:44	**Log-Likelihood:**	-276.41			
Converged:	True	**LL-Null:**	-282.63			
Covariance Type:	nonrobust	**LLR p-value:**	0.006041			
	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
AI Comment with Negative Score	-2.0365	1.091	-1.867	0.062	-4.175	0.102
AI Comment with Positive Score	-1.8135	1.071	-1.694	0.09	-3.912	0.285
Child Comment also Discussing AI	-0.3025	0.147	-2.056	0.04	-0.591	-0.014
Top Level Comment Indicates AI?	1.9209	1.069	1.796	0.072	-0.175	4.017

For the purposes of this work, all four metrics will be presented, but the most valuable characterization of the model's performance is precision. When running the model against social media posts in practice, we cannot expect to catch every AI-generated submission, but for this methodology to provide the most utility, the positive predictions produced should be reliably correct. The three viability criteria listed in Section 3.2 are evaluated using the *statsmodels* Python package [30].

Results. When selecting from the potential logistic regression predictors, the best fit regression model was created using the "AI Comment with Negative Score," "AI Comment with Positive Score," "Child Comment also Discussing AI," and "Top Level Comment Indicates AI?" independent variables. The P value for each coefficient, as seen in Table 1, are all less than 0.1 with a maximum value of 0.09 and a minimum value of 0.04 in "AI Comment with Positive Score" and "Child Comment also Discussing AI," respectively. Due to the low P values presented, we can claim that these independent variables have an impact on the outcome with greater than 90% significance. The overall model's P value is also less than 0.1 at approximately 0.006, also indicating greater than 90% significance.

However, the logistic regression model is not a complete success. The confusion matrix shown in Table 2a highlights the model's prediction results. Of the 272 submissions in the test split, 92 were true negatives, 63 were true positives, 45 were false positives, and 72 were false negatives. The model's performance metrics are as follows:

Table 2: Confusion matrices for the base interaction-based pipeline (left) and the LLM-enabled pipeline (right).

	Pred. Neg	*Pred. Pos*
Actual Neg	**92**	45
Actual Pos	72	**63**

(a) Base model.

	Pred. Neg	*Pred. Pos*
Actual Neg	**115**	22
Actual Pos	90	**45**

(b) LLM-enabled model.

- Accuracy: 56.99%
- Precision: 58.33%
- Recall: 46.66%
- F1 Score: 51.85%

The appendix (Section A) highlights a few real outcomes generated by the regression model, their feature vectors, and the classification results to best illustrate the model's detection capability.

3.4 LLM-Enabled Preprocessing and Post-Engine Analysis Pipeline

In observing the outcomes of the standard, interaction-based detection pipeline, it is apparent that there are flaws in the existing system, resulting in an accuracy not much better than a coin-flip. By using large language models to supplement the detection algorithm, false positives can be reduced and performance, the most significant performance metric in this scope, can be improved.

Post-Engine Pipeline Design Goals. We designed the LLM-enabled detection model to improve upon the base model in the following criteria, in no particular order. By accomplishing these design goals, the detection model's overall efficiency and performance can be improved.

- **Goal 1:** ***Preprocessing Out Irrelevant Submissions.*** Closer examination of the base model's erroneous results shows that many of the false positives are for posts that seem contextually improbable as targets for synthetic-content generation. By using a large language model to identify content themes that would be illogical for a deepfake, preprocessing and filtering them out before running the submission through the base model would reduce potential for them to be attributed to a false positive result.
- **Goal 2:** ***Gauging Users' Conviction in AI Accusations.*** This work recognizes that user interactions accusing a submission of having AI-generated origins are made with varying levels of conviction, depending on the users' understanding of synthetic content. There is a difference in a user claiming to be uncertain in their judgment and another stating matter-of-factly that the content is AI-generated and providing reasoning of why. These discrepancies in conviction strength should be treated accordingly.

– **Goal 3:** ***Subject Extraction to Enable Theme-Based Detection Strategy.*** This work proposes an alternative methodology to implement a 'known falsehoods' database recording previously deepfaked contents' themes in an effort to explore the hypothesis that previously AI-generated themes are likely targets for future deceptions. This idea is explored further in Section 4.

Performance Enhancements. To build the LLM-enabled preprocessing and post-engine analysis pipeline, this project uses Meta's Llama 3.2:3b model [1] to implement subject extraction and conviction mapping. The LLM-enabled model presents the Llama 3.2:3b model with information about each submission and provides it with a series of prompts to extract information about the subject and its relevance.

First, the model is given the submission's title and body text, separated with a unique token, and told the following: "Your job is to take a 1 or two pieces of text, separated by the token "&*&%". Looking at these texts, I need you to tell me what the subject is. Please do not say anything more, only the subject you identified. If you cannot determine the subject, please only say "unknown."" The chosen subject is then passed back to the model as it is instructed to "take a piece of text that indicates a subject or topic matter. Looking at this text, tell me if the subject is of interest. Examples of subjects of interest are paintings, or music. Examples of subjects not of interest are someone's name, political themes, and anything that is not creative. Please do not say anything more, only "Yes" or "No". You must make a decision." Depending on the result of this complete relevancy decision, a post can either be marked as human-generated right away or proceed through the rest of the pipeline as usual.

The confusion matrix generated after implementing Design Goals 1 and 2 can be seen in Table 2b. For purposes of drawing fair comparisons between the LLM-enabled detection model and the base detection model, the same 60:40 train/test split seed divided the labeled data. The logistic regression model remained unchanged for both trials.

Of the 272 submissions in the test split, 115 were true negatives, 45 were true positives, 22 were false positives, and 90 were false negatives. The LLM-enabled model's performance metrics are as follows:

– Accuracy: 58.82%
– Precision: 67.16%
– Recall: 33.33%
– F1 Score: 44.55%

4 Discussion

As generative AI continues to rapidly develop, social media platforms present a playground for bad actors to distill their deceitful, deepfaked content. Gone unchecked, misinformation, fake news, and distrust can be spread over the internet like a wildfire. Current solutions attempt to target the generated content

itself and fail to leverage any of the unique resources that social media platforms make available. This work developed an interaction-based solution and explored user insights as a way to detect AI-generated content on social media.

The results presented in this project show that our novel methodology underperforms relative to current content-based solutions, including Baraheem & Nguyen's (2023) generalized GAN-generated image detection model and Mitra et al.'s (2021) key video frame extraction model. These methodologies have detection accuracies of 89% and 92%, respectively, while our LLM-enabled interaction-based pipeline achieves an accuracy of approximately 59% and a precision around 67%. While this performance gap is significant, it is important to consider the broader implications of utilizing human-generated sentiments on social media as a basis for detecting AI-generated content.

Unlike content-based detection methodologies, trained to identify implicit artifacts attributed to upsampling operations common in generative-AI models or to evaluate a text's intrinsic structure and comparing against common patterns in human writing, this work relies almost exclusively on human judgment to identify and confirm AI-involvement. As in all things, humans are prone to errors. Interestingly, current research in adjacent fields corroborates this sentiment.

Liu et al. (2024) [24] looks into the ability for automated AI detection tools, such as GPTZero and Turnitin, and human reviewers to make assessments about a document's authenticity. Their study gave four reviewers, made up of two professors and two college students, a series of academic publications in their fields of study and were instructed to determine if they believed each document to be human- or AI-generated. While the reviewers did well in classifying AI-rephrased articles, they incorrectly mislabeled 18% of human-written articles as AI-generated. Even though they are considered to be highly familiar with, and even experts in the field they were critiquing work in, they were still fairly prone to misattribute texts the lengths of full research papers. When making their decisions, reviewers were required to include a reasoning as to why they believed the document to be AI-rephrased. Nearly 55% of the time, the reviewers stated "incoherence" or "grammatical errors" were the primary indicators. In a social media setting where the conversational etiquette is far more lax than in an academic setting, incoherence, grammatical errors, and brief texts to draw conclusions from are commonplace.

Cooke et al. (2024) [14] tasked 1,276 randomly-selected participants with classifying human- and AI-generated content across different media types including audiovisual, audio-only, video-only, and images. The mean detection performance across all media types was only 51.2%. The most accurately classified media type was audiovisual and the least were images at 54.5% and 49.4%, respectively. Participants self-reported being previously unfamiliar with synthetic media and those who were, either semi-familiar or highly familiar. There was no significant difference in detection performance between the three groups.

The work done by Chein et al. (2024) [11] most closely parallels that done in this work, exploring the way AI-generated content can be unknowingly spread over social media without users discerning it as AI. 194 participants were pre-

sented with online general interest news story headlines and online scientific news story headlines, similar to the ones that could logically be shared online. A random subset from 48 human-generated texts and another 48 matching AI-generated texts created with ChatGPT were shown to each participant for classification. On average, participants were 57% accurate in identifying the origin of human- and AI-generated texts. When separated by source, participants correctly attributed human-generated content to human authorship accurately 61% of the time on average, and the AI-generated content to AI authorship 53% of the time.

Unlike these studies, where the participants were explicitly tasked with classifying texts as human- or AI-generated, the performance results in our work are derived from social media users' ability to voluntarily identify a post as AI generated. Interestingly, this work reinforces humans' inability to accurately identify AI-generated content, and the accuracy of commenters against our data is no different from the results Liu et al., Cooke et al., and Chein et al. arrived to in their works.

There are several limitations worth noting. Namely, all content used throughout this project came from Reddit. Though the methodology was designed to be platform-agnostic, testing against data on other social media platforms could expose weaknesses in our methodology or confirm the model's efficacy in a new environment and engagement structure. The other is that the detection model was tested thoroughly against our labeled dataset, not unseen submissions in the wild. This decision was deliberately made to counteract the heavy skew in human-generated content over AI-generated content on Reddit. Focusing on a social media platform with a higher density of AI-generated content, such as Quora or Medium [33], could generate more viable data without needing to artificially alter the training dataset's sample population with downsampling.

Future research into a secondary, theme-based detection strategy building off the interaction-based model, exploring how a 'known falsehoods' dataset, recording previously deepfaked content themes, could be implemented to support its human- or AI-generated attribution. It may also be beneficial to investigate the relationship between previously identified deepfake subjects and the prevalence of similar deepfakes in the future, noting any new patterns that emerge.

Although the interaction-based detection approach falls short of existing detection benchmarks in terms of raw performance, these findings reinforce the need for a generalizable detection tool. One that takes the decision-making out of human hands as generative AI continues to improve and our ability in detecting AI-generated material approaches the odds of a coin-flip.

Acknowledgments. Thanks to the CyberCorps®: Scholarship For Service cohort at Worcester Polytechnic Institute for their support throughout this research.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

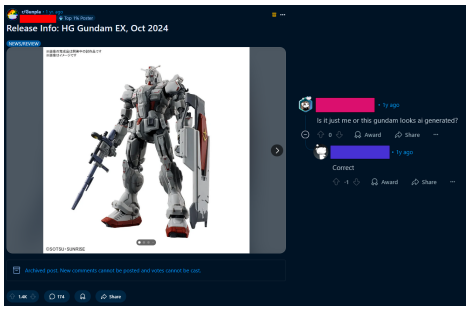
References

1. llama3.2, https://ollama.com/llama3.2
2. PRAW 7.7.1 documentation, https://praw.readthedocs.io/en/stable/
3. Aguilera, A., Quinteros, P., Dongo, I., Cardinale, Y.: CrediBot: Applying Bot Detection for Credibility Analysis on Twitter. IEEE Access **11**, 108365–108385 (2023). https://doi.org/10.1109/ACCESS.2023.3320687, https://ieeexplore.ieee.org/document/10267924, conference Name: IEEE Access
4. Alharbi, A., Dong, H., Yi, X., Tari, Z., Khalil, I.: Social Media Identity Deception Detection: A Survey. ACM Comput. Surv. **54**(3), 69:1–69:35 (Apr 2021). https://doi.org/10.1145/3446372, https://dl.acm.org/doi/10.1145/3446372
5. Arık, K.: Social Media Content Review of Popular MMORPG Games: Reddit Comment Scraping and Sentiment Analysis. Journal of Emerging Computer Technologies **2**(1), 13–21 (Jul 2022), https://dergipark.org.tr/en/pub/ject/issue/68031/1103252, number: 1 Publisher: İzmir Academy Association
6. Baraheem, S.S., Nguyen, T.V.: AI vs. AI: Can AI Detect AI-Generated Images? Journal of Imaging **9**(10), 199 (Oct 2023). https://doi.org/10.3390/jimaging9100199, https://www.mdpi.com/2313-433X/9/10/199, number: 10 Publisher: Multidisciplinary Digital Publishing Institute
7. Batrinca, B., Treleaven, P.C.: Social media analytics: a survey of techniques, tools and platforms. AI & SOCIETY **30**(1), 89–116 (Feb 2015). https://doi.org/10.1007/s00146-014-0549-4, https://doi.org/10.1007/s00146-014-0549-4
8. Baydogan, C., Alatas*, B.: Sentiment Analysis in Social Networks Using Social Spider Optimization Algorithm. Tehnički vjesnik **28**(6), 1943–1951 (Nov 2021). https://doi.org/10.17559/TV-20200614172445, https://hrcak.srce.hr/clanak/383559, publisher: Sveučilište u Slavonskom Brodu, Stojarski fakultet
9. Bick, A., Blandin, A., Deming, D.J.: The Rapid Adoption of Generative AI (Sep 2024). https://doi.org/10.3386/w32966, https://www.nber.org/papers/w32966
10. Chatzakou, D., Soler-Company, J., Tsikrika, T., Wanner, L., Vrochidis, S., Kompatsiaris, I.: User Identity Linkage in Social Media Using Linguistic and Social Interaction Features. In: Proceedings of the 12th ACM Conference on Web Science. pp. 295–304. WebSci '20, Association for Computing Machinery, New York, NY, USA (Jul 2020). https://doi.org/10.1145/3394231.3397920, https://dl.acm.org/doi/10.1145/3394231.3397920
11. Chein, J., Martinez, S., Barone, A.: Can human intelligence safeguard against artificial intelligence? Exploring individual differences in the discernment of human from AI texts (Apr 2024). https://doi.org/10.21203/rs.3.rs-4277893/v1, https://www.researchsquare.com/article/rs-4277893/v1, iSSN: 2693-5015
12. Christanti, M.V., Walda, Sutrisno, T.: Comments Scraping Application For Review Youtube Content. IOP Conference Series: Materials Science and Engineering **852**(1), 012167 (Jul 2020). https://doi.org/10.1088/1757-899X/852/1/012167, https://dx.doi.org/10.1088/1757-899X/852/1/012167, publisher: IOP Publishing
13. Chung, W., Zeng, D.: Dissecting emotion and user influence in social media communities: An interaction modeling approach. Information & Management **57**(1), 103108 (Jan 2020). https://doi.org/10.1016/j.im.2018.09.008, https://www.sciencedirect.com/science/article/pii/S0378720617309229
14. Cooke, D., Edwards, A., Barkoff, S., Kelly, K.: As Good As A Coin Toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli

(Apr 2024). https://doi.org/10.48550/arXiv.2403.16760, http://arxiv.org/abs/2403.16760, arXiv:2403.16760 [cs]
15. Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., Verdoliva, L.: Raising the Bar of AI-generated Image Detection with CLIP. pp. 4356–4366 (2024), https://openaccess.thecvf.com/content/CVPR2024W/WMF/html/Cozzolino_Raising_the_Bar_of_AI-generated_Image_Detection_with_CLIP_CVPRW_2024_paper.html
16. Dev, H., Ali, M.E., Hashem, T.: User Interaction Based Community Detection in Online Social Networks. In: Bhowmick, S.S., Dyreson, C.E., Jensen, C.S., Lee, M.L., Muliantara, A., Thalheim, B. (eds.) Database Systems for Advanced Applications. pp. 296–310. Springer International Publishing, Cham (2014). https://doi.org/10.1007/978-3-319-05813-9_20
17. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The DeepFake Detection Challenge (DFDC) Dataset (Oct 2020). https://doi.org/10.48550/arXiv.2006.07397, http://arxiv.org/abs/2006.07397, arXiv:2006.07397 [cs]
18. Epstein, D.C., Jain, I., Wang, O., Zhang, R.: Online Detection of AI-Generated Images. pp. 382–392 (2023), https://openaccess.thecvf.com/content/ICCV2023W/DFAD/html/Epstein_Online_Detection_of_AI-Generated_Images__ICCVW_2023_paper.html
19. Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M., Fdez-Riverola, F.: Web scraping technologies in an API world. Briefings in Bioinformatics **15**(5), 788–797 (2013). https://doi.org/10.1093/bib/bbt026
20. Golzadeh, M., Decan, A., Legay, D., Mens, T.: A ground-truth dataset and classification model for detecting bots in GitHub issue and PR comments. The Journal of systems and software **175**, 110911– (2021). https://doi.org/10.1016/j.jss.2021.110911, publisher: Elsevier Inc
21. Guo, Q., Xie, H., Li, Y., Ma, W., Zhang, C.: Social Bots Detection via Fusing BERT and Graph Convolutional Networks. Symmetry (Basel) **14**(1), 30– (2022). https://doi.org/10.3390/sym14010030, number: 1 Place: Basel Publisher: MDPI AG
22. Juba, B., Le, H.S.: Precision-Recall versus Accuracy and the Role of Large Data Sets. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 4039–4048 (Jul 2019). https://doi.org/10.1609/aaai.v33i01.33014039, https://ojs.aaai.org/index.php/AAAI/article/view/5193, number: 01
23. Kang, A.R., Kim, H.K., Woo, J.: Chatting Pattern Based Game BOT Detection: Do They Talk Like Us? KSII transactions on Internet and information systems **6**(11), 2866–2879 (2012)
24. Liu, J.Q.J., Hui, K.T.K., Al Zoubi, F., Zhou, Z.Z.X., Samartzis, D., Yu, C.C.H., Chang, J.R., Wong, A.Y.L.: The great detectives: humans versus AI detectors in catching large language model-generated medical writing. International Journal for Educational Integrity **20**(1), 1–14 (Dec 2024). https://doi.org/10.1007/s40979-024-00155-6, https://edintegrity.biomedcentral.com/articles/10.1007/s40979-024-00155-6, number: 1 Publisher: BioMed Central
25. Mathys, M., Willi, M., Meier, R.: Synthetic Photography Detection: A Visual Guidance for Identifying Synthetic Images Created by AI (Aug 2024). https://doi.org/10.48550/arXiv.2408.06398, http://arxiv.org/abs/2408.06398, arXiv:2408.06398 [cs]
26. Mitra, A., Mohanty, S., Corcoran, P., Kougianos, E.: A Machine Learning Based Approach for Deepfake Detection in Social Media Through Key Video Frame

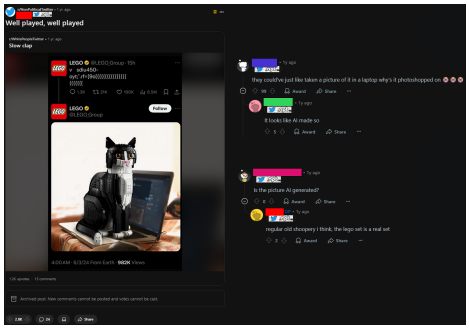
Extraction. SN Computer Science **2** (Apr 2021). https://doi.org/10.1007/s42979-021-00495-x
27. OpenAI: ChatGPT
28. OpenAI: DALL·E 3, https://openai.com/index/dall-e-3/
29. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces (Mar 2018). https://doi.org/10.48550/arXiv.1803.09179, http://arxiv.org/abs/1803.09179, arXiv:1803.09179 [cs]
30. Seabold, S., Perktold, J.: statsmodels: Econometric and statistical modeling with python. In: 9th Python in Science Conference (2010)
31. Solanki, C.: "This Has Been Written by a Bot": A Bot Detection Study of the SubsimulatorGPT2 Subreddit. In: Chakrabarti, A., Singh, V. (eds.) Design in the Era of Industry 4.0, Volume 1. pp. 1183–1190. Springer Nature, Singapore (2023). https://doi.org/10.1007/978-981-99-0293-4_95
32. stuck_in_the_matrix, Watchful1, RaiderBDev: Reddit comments/submissions 2005-06 to 2024-06 (2024), https://academictorrents.com/details/20520c420c6c846f555523babc8c059e9daa8fc5
33. Sun, Z., Zhang, Z., Shen, X., Zhang, Z., Liu, Y., Backes, M., Zhang, Y., He, X.: Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media (Dec 2024). https://doi.org/10.48550/arXiv.2412.18148, http://arxiv.org/abs/2412.18148, arXiv:2412.18148 [cs]
34. Wang, P., Li, L., Ren, K., Jiang, B., Zhang, D., Qiu, X.: SeqXGPT: Sentence-Level AI-Generated Text Detection (Dec 2023). https://doi.org/10.48550/arXiv.2310.08903, http://arxiv.org/abs/2310.08903, arXiv:2310.08903 [cs]
35. Willman, J.: Overview of PyQt5. In: Willman, J. (ed.) Modern PyQt: Create GUI Applications for Project Management, Computer Vision, and Data Analysis, pp. 1–42. Apress, Berkeley, CA (2021). https://doi.org/10.1007/978-1-4842-6603-8_1, https://doi.org/10.1007/978-1-4842-6603-8_1
36. Yu, C.M., Chen, K.C., Chang, C.T., Ti, Y.W.: SegNet: a network for detecting deepfake facial videos. Multimedia Systems **28**(3), 793–814 (Jun 2022). https://doi.org/10.1007/s00530-021-00876-5, https://doi.org/10.1007/s00530-021-00876-5

A Case Studies in Classification Performance



Feature	Value
Art Community	0
Sensitive Topic	1
Active Downvoted	0
Top-Level Cmt AI	1
Author Confirms AI	0
AI Cmt Positive	0
AI Cmt Negative	1
AI Child	0

Fig. 6: False positive prediction (FP1) due to improper prediction weight favoring commenters over nonvocal users disagreeing with the statements claiming AI; feature vector at right.

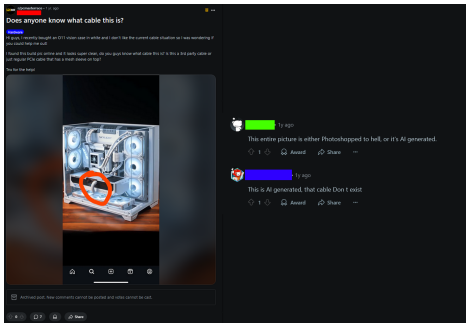


Feature	Value
Art Community	0
Sensitive Topic	1
Active Downvoted	0
Top-Level Cmt AI	1
Author Confirms AI	0
AI Cmt Positive	0
AI Cmt Negative	1
AI Child	0

Fig. 7: False positive prediction (FP2) caused by false positive result in sensitive topic classification.

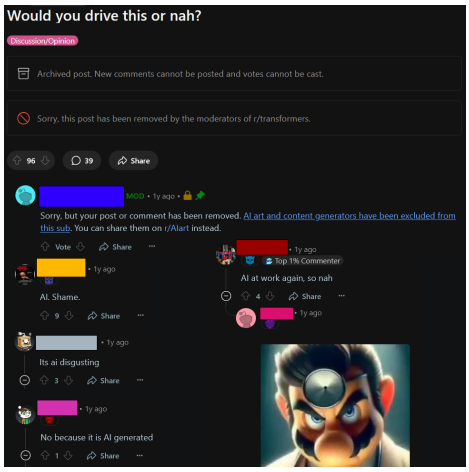
Feature	Value
Art Community	0
Sensitive Topic	0
Active Downvoted	0
Top-Level Cmt AI	1
Author Confirms AI	0
AI Cmt Positive	1
AI Cmt Negative	0
AI Child	0

Fig. 8: False positive (FP3): submission falsely attributed as being AI-generated. The detection model picked up on a generic moderator-served comment disclosing the community's rules.



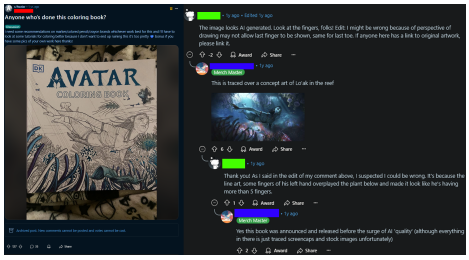
Feature	Value
Art Community	0
Sensitive Topic	0
Active Downvoted	0
Top-Level Cmt AI	3
Author Confirms AI	0
AI Cmt Positive	3
AI Cmt Negative	0
AI Child	0

Fig. 9: True positive (TP1) yielded after the model accurately identified positively received comments claiming AI.



Feature	Value
Art Community	0
Sensitive Topic	0
Active Downvoted	0
Top-Level Cmt AI	5
Author Confirms AI	0
AI Cmt Positive	5
AI Cmt Negative	0
AI Child	0

Fig. 10: True positive (TP2): submission correctly predicted to be AI-generated from a high frequency of well-received comments accusing the post of being AI-generated.



Feature	Value
Art Community	0
Sensitive Topic	0
Active Downvoted	0
Top-Level Cmt AI	1
Author Confirms AI	0
AI Cmt Positive	0
AI Cmt Negative	1
AI Child	0

Fig. 11: True negative (TN1): submission correctly identified as human-generated. Though one commenter believed the post AI-generated, disagreement and negative voting dismissed the claim.