



ETHICAL AND BIAS AWARE FAKE NEWS DETECTION SYSTEM



A PROJECT WORK

Submitted by

ARAVIND G (71812111006)

LAKSHAN V K (71812111023)

SANJAY R (71812111043)

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

SRI RAMAKRISHNA ENGINEERING COLLEGE COIMBATORE

ANNA UNIVERSITY : CHENNAI 600 025

MAY 2025



**SRI RAMAKRISHNA ENGINEERING COLLEGE
COIMBATORE**



ANNA UNIVERSITY : CHENNAI 600025

BONAFIDE CERTIFICATE

Certified that this major project Report “**ETHICAL AND BIAS AWARE FAKE NEWS DETECTION SYSTEM**” is the bonafide work of **ARAVIND G (71812111006), LAKSHAN V K (71812111023), SANJAY R (71812111043)** who carried out the **20AD279 - Project work** under my supervision.

SIGNATURE

SUPERVISOR

Dr. B. Suganya
Assistant Professor
Department of Artificial Intelligence
and Data Science,
Sri Ramakrishna Engineering College,
Coimbatore-641022

SIGNATURE

HEAD OF THE DEPARTMENT

Dr. V. Karpagam
Professor and Head,
Department of Artificial Intelligence
and Data Science,
Sri Ramakrishna Engineering College,
Coimbatore-641022

Submitted for Project work Viva Voce Examination held on _____

INTERNAL EXAMINER

EXTERNAL EXAMINER

5.2.1 SYSTEM ARCHITECTURE

The architecture specifies an explainable and ethical AI-powered automated false news detection. The approach starts with data collection, where the dataset is diversified to include news reports of different sources and subjects. Data preparation and feature engineering follow, transforming the raw text content to structured input for model training. The primary classification is performed by a RoBERTa-based transformer model that reads linguistic patterns to determine if content is authentic or not. After classification, an explanation module is triggered through LIME and SHAP to determine significant words or phrases that resulted in the prediction. These highlighted features offer transparency and enable users to comprehend why a news article was labelled fake or real. In found potential bias cases, mitigation measures and fairness-aware algorithms are activated to restrain biased outcomes. Explanation data, found biases, and classification results are saved to a centralized database. Lastly, the outcome may be displayed in a real-time dashboard, wherein real-time feedback is given.

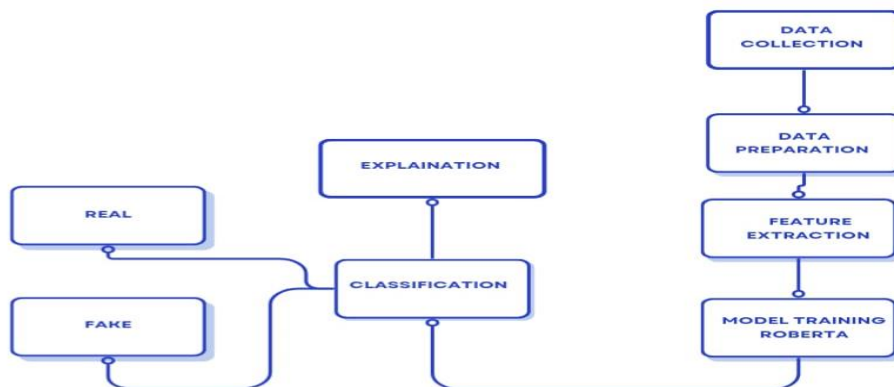


Fig 5.1 System Architecture

responsive and professional layout. JavaScript, through renders interactive visualizations and handles real-time API calls. Flask's RESTful endpoints provide classifications, explanations, and warnings and integrate fact-checking.

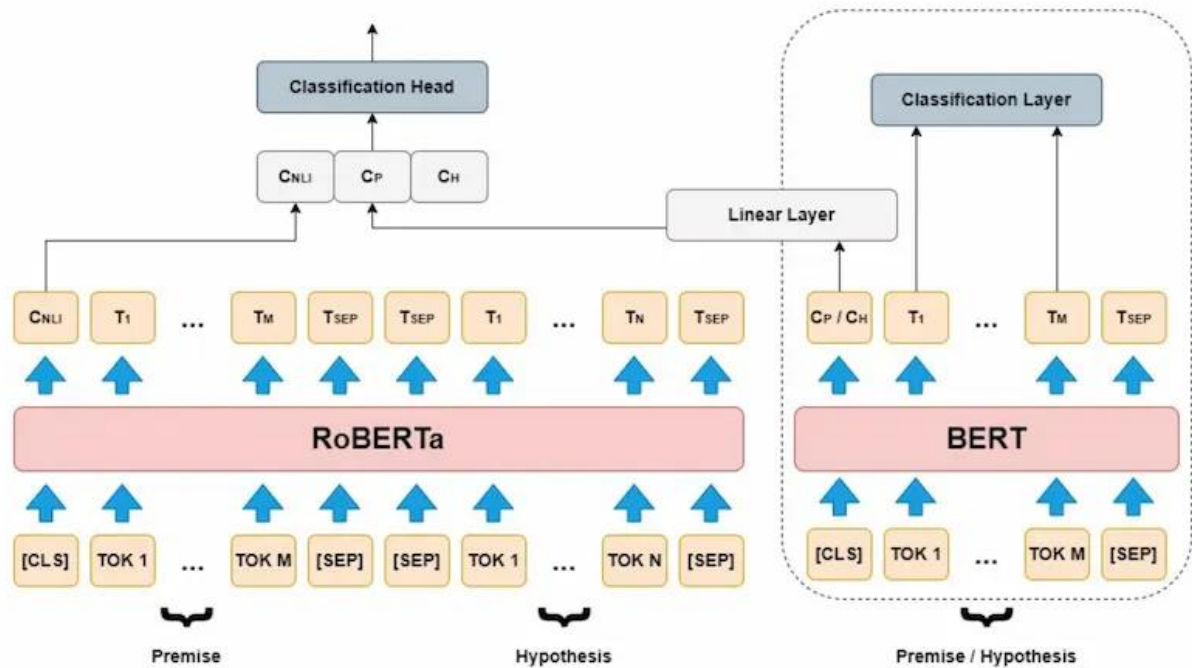


Fig 5.2 RoBERTa Architecture

5.3.4 MODEL TRAINING

The fine-tuning of RoBERTa lies at the core of the effectiveness of the system, aimed at attaining high precision while maintaining Responsible AI principles. The model is developed from a heterogeneous set of twenty thousand news articles, sixty percent being labelled as authentic and forty percent being labelled as fabricated, collected from news channels, fact-checking websites, and social media. The dataset has been split into fourteen thousand training instances, three thousand validation instances, and three thousand testing instances across topics like politics, health, and technology for wide usability.

risks, and cross-device compatibility is tested. Such integration is an expression of Responsible AI, giving a just, transparent, and people-oriented platform empowering effective misinformation mitigation with ethical integrity.

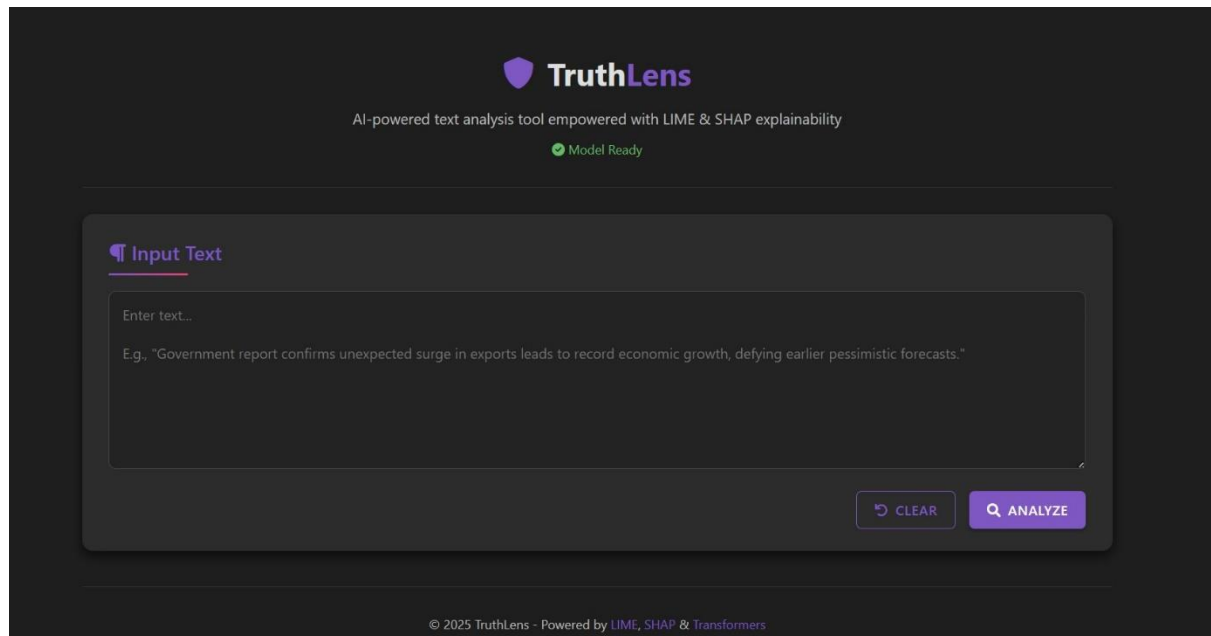


Fig 5.5 User Interface Home Page

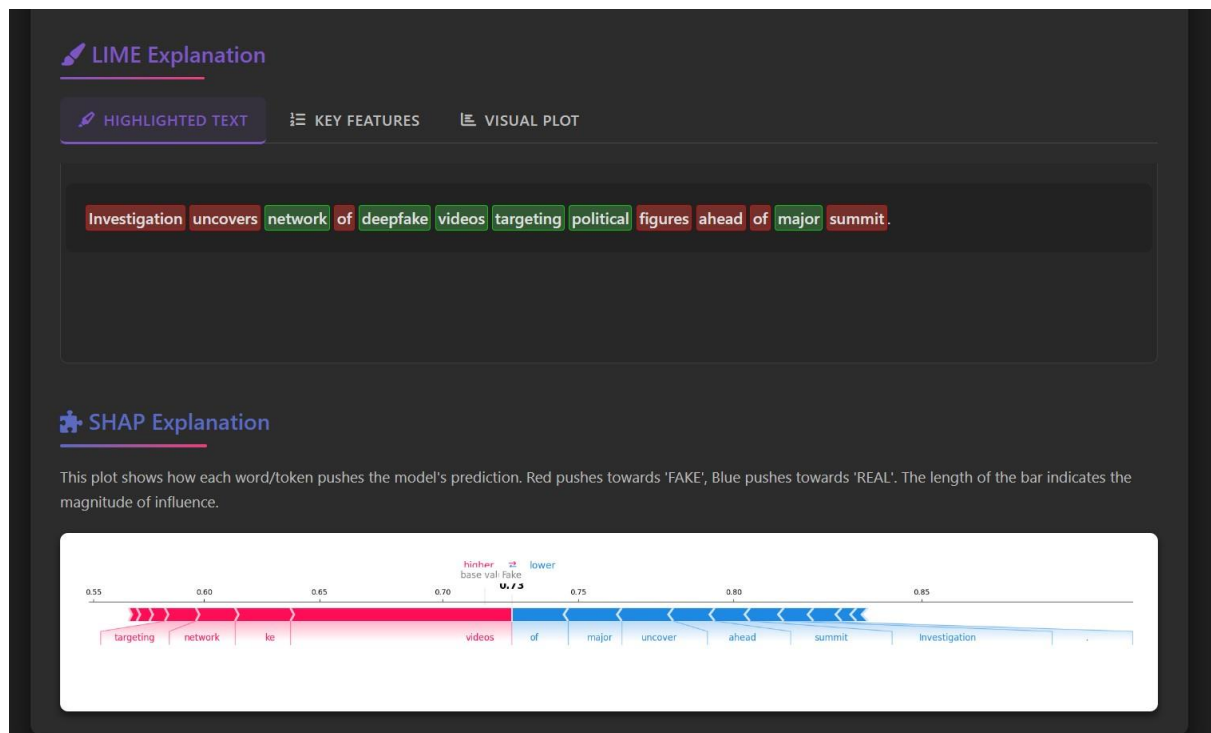


Fig 5.6 User Interface Explanation of LIME & SHAP

CHAPTER 6

RESULT AND DISCUSSION

6.1 EXPERIMENTAL RESULTS

The Ethical and Bias-Aware Fake News Detection System is proposed to detect fake news in a way that preserves transparency and fairness. To test the performance, an extensive study has been carried using a balanced sample of 3,000 articles (1,500 real and 1,500 fake), randomly chosen from a larger collection of 20,000 news stories, of which 60 percent were real and 40 percent were fake. The articles covered wide-ranging topics including health and politics. On utilizing five-fold cross-validation to bolster the reliability of the assessment and to avoid overfitting. The base model, RoBERTa was trained for two epochs with a batch size of 32, and early stopping after 500 uninterrupted steps without convergence improvement. For the sake of comparison, testing a baseline LSTM model under the same training settings.

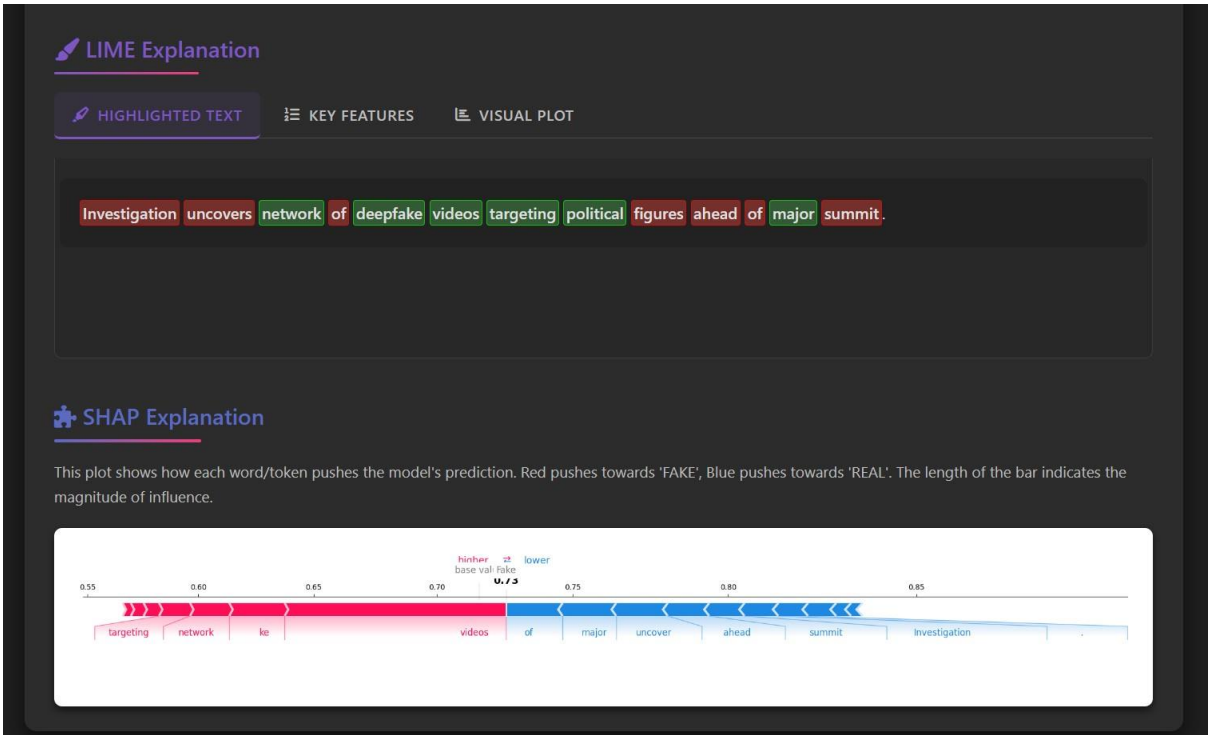
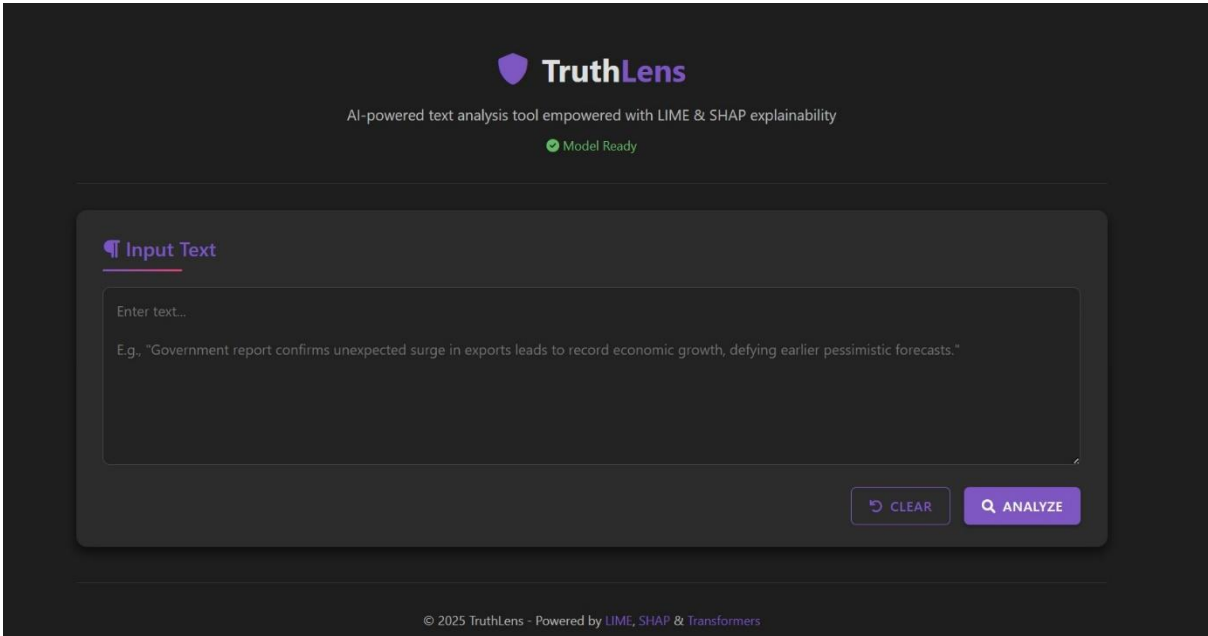
In order to make model explainability more accessible and improve user trust, LIME and SHAP have been coupled in a Flask web interface. Fifty participants used the system through the web interface and reported back about the model's predictability transparency and interpretability



Fig 6.1 LIME Result

ANNEXURE II

SCREENSHOTS



ANNEXURE III

BASE PAPER

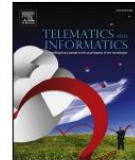
Telematics and Informatics 91 (2024) 102135



Contents lists available at ScienceDirect

Telematics and Informatics

journal homepage: www.elsevier.com/locate/tele



EXplainable Artificial Intelligence (XAI) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives

Ching-Hua Chuan^a, Ruoyu Sun^b, Shiyun Tian^c, Wan-Hsiu Sunny Tsai^{d,*}

^a Department of Interactive Media, University of Miami, 5100 Brunson Drive, Coral Gables, FL 33134, USA

^b Department of Advertising and Public Relations, University of Georgia, 120 Hooper St, Athens, GA 30602-3018, USA

^c Department of Marketing, Sacred Heart University, 5151 Park Avenue, Fairfield, CT 06825, USA

^d Department of Strategic Communication, University of Miami, 5100 Brunson Drive, Coral Gables, FL 33134, USA

ARTICLE INFO

Keywords:

eXplainable artificial intelligence (XAI)
Algorithmic bias
Fairness
Inclusiveness
Imposed users

ABSTRACT

This study explored the potential of eXplainable Artificial Intelligence (XAI) in raising user awareness of algorithmic bias. This study examined the popular “explanation by example” XAI approach, where users receive explanatory examples resembling their input. As this XAI approach allows users to gauge the congruence between these examples and their circumstances, perceived incongruence then evokes perceptions of unfairness and exclusion, prompting users not to put blind trust in the system and raising awareness of algorithmic bias stemming from non-inclusive datasets. The results further highlight the moderating role of users' prior experience with discrimination.

1. Introduction

Artificial intelligence (AI) has been widely recognized as a double-edged sword, improving every aspect of our lives and at the same time posing critical threats to societal well-being, particularly because AI is rapidly being integrated into decision-making processes, from resume screening for employment decisions to financial loan approval. However, AI biases based on race, gender, sexual orientation, and other identity factors have been increasingly observed in the outputs of AI systems (Kordzadeh and Ghasemaghaei, 2022). Algorithmic bias can be found in common, everyday applications, such as search engines' autocomplete predictions that produce more negative biases for female, Black, and homosexuality-related prefixes (Lin et al., 2023), and in domain-specific tools, such as sentencing and risk assessment software in criminal justice that discriminates against Black defendants (Angwin et al., 2016). As AI algorithms are deployed via industry-wide mass automation, disparities and inequities are exacerbated at a much greater scale and speed, generating growing concerns on algorithmic bias (Chuan et al., 2023).

Algorithmic bias is defined as “systematic and structured errors in an artificial intelligence system that generate unfair results and inequalities” (Shin and Shin, 2023, p. 90). Critically, due to the “black-box” nature of complicated AI systems, even AI developers cannot reliably explain why a specific (potentially biased) decision is produced. Therefore, facilitating acute awareness and recognition of potential algorithmic biases among various stakeholders within the AI ecosystem is a crucial first step for mitigating the

* Corresponding author.

E-mail addresses: c.chuan@miami.edu (C.-H. Chuan), rsun@uga.edu (R. Sun), tians@sacredheart.edu (S. Tian), wanhsui@miami.edu (W.-H.S. Tsai).

<https://doi.org/10.1016/j.tele.2024.102135>

Received 8 February 2024; Received in revised form 29 April 2024; Accepted 9 May 2024

Available online 10 May 2024

0736-5853/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

ANNEXURE IV

PAPER PUBLICATION ACCEPTANCE

Dear Author,
Greetings and best wishes of the day !!

Our International Conference has accepted your Research Paper entitled "ETHICAL AND BIAS AWARE FAKE NEWS DETECTION SYSTEM" with Paper ID **IST-BDE-PNDC-190525-6964**

Note : 8th May is the last date of Registration.
There are no additional charges for Additional Certificates for Additional Authors and for Publication.
To avoid Bank Charges you can make a Bank Transfer through any UPI Application by using our Account Number and IFSC Code.

Online payment link <http://paymentnow.in/>

OR

Bank Details

EVER LIFE HEALTH PRIVATE LIMITED

Ac/no.: 50200052618798

HDFC Bank, Nayapalli, Bhubaneswar, Odisha, India

IFSC: HDFC0000640

SWIFT CODE: HDFCINBB

Only after your registration has been confirmed will you receive a Formal acceptance letter and Conference Schedule.





ANNEXURE V

PLAGIARISM REPORT




7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

-  **33 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
-  **0 Missing Quotations 0%**
Matches that are still very similar to source material
-  **6 Missing Citation 1%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5%  Internet sources
- 5%  Publications
- 0%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

*% detected as AI

AI detection includes the possibility of false positives. Although some text in this submission is likely AI generated, scores below the 20% threshold are not surfaced because they have a higher likelihood of false positives.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Disclaimer

Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?

The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

