SRI RAMAKRISHNA ENGINEERING COLLEGE
[Educational Service: SNR Sons Charitable Trust]
[Autonomous Institution, Reaccredited by NAAC with 'A+' Grade]
[Approved by AICTE and Permanently Affiliated to Anna University, Chennai]
[ISO 9001:2015 Certified and all eligible programmes Accredited by NBA]
Vattamalaipalayam, N.G.G.O. Colony Post, Coimbatore – 641 022.

DEPARTMENT OF
ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

20AD279 - Project work

Ethical and Bias-Aware Fake News Detection System

TEAM MEMBERS

ARAVIND G (2111006)
LAKSHAN V K (2111023)
SANJAY R (2111043)

MENTOR

Dr. B. Suganya
Assistant Professor/AI&DS

Agenda

- Domain
- Problem Statement
- Objective
- Literature Survey
- Existing System
- Proposed System
- Block Diagram
- Modules
- Challenges Faced and Solutions
- Models Comparison
- UI
- Outputs
- Paper Acceptance
- Conclusion
- References
- Future Scope

DOMAIN

DEEP LEARNING

- Deep learning is a subset of machine learning that uses artificial neural networks with multiple layers to learn patterns from large datasets.

- It excels in handling complex data like images, text, ad audio. Deep learning powers applications such as image recognition, natural language processing, and autonomous systems, requiring significant data and computational resources.

EXPLAINABLE AI

- Explainable AI aims to make AI models transparent and interpretable by explaining their decision-making processes.

- It enhances trust, accountability, and fairness in AI systems. Techniques like LIME, SHAP, and saliency maps help users understand model predictions, ensuring ethical use in critical fields such as healthcare, finance, and law.

PROBLEM **STATEMENT**

- The challenge lies in the complexity of accurately identifying fake news, given the dynamic nature of misinformation, intricate language patterns and potential biases in training data.

- The proliferation of fake news on digital platforms undermines public trust and decision-making.

- Existing detection systems often function as "black boxes," lacking transparency and fairness.

- These models may inadvertently introduce biases, resulting in unfair predictions that disproportionately affect certain groups sources or topics further perpetuating misinformation.

OBJECTIVE

- To develop an accurate fake news detection system, this project utilizes deep learning models, tackling the complexities of evolving misinformation and biased data.

- To enhance model transparency, explainable AI techniques are incorporated, ensuring users can understand the reasoning behind predictions.

- To promote fairness, the project focuses on mitigating biases in the detection process.

LITERATURE **SURVEY**

S.NO	TITLE	YEAR & PUBLICATION	METHODOLOGY USED	LIMITATIONS
1	Explainable Fake News Detection With Large Language Model via Defense Among Competing Wisdom	2024 (ACM Web Conference)	• LLMs for evidence extraction • Defense-based inference • Prompt-based justifications	• Majority bias • Misleading justifications • Over-reliance on AI
2	A Review on Fake News Detection using Deep Learning Methods	2024, International Journal of Scientific Research in Computer Science, Engineering and Information Technology	• CNN • RNN • LSTM	• Limited adaptability • Poor Explainability • Incomplete exploration

LITERATURE **SURVEY**

S.NO	TITLE	YEAR & PUBLICATION	METHODOLOGY USED	LIMITATIONS
3	When Explainability Turns into a Threat – Using xAI to Fool a Fake News Detection Method	2024, Computers & Security (Elsevier)	• SHAP • Adversarial attacks • Explainability vulnerabilities	• Explainability exploitation • Security vulnerabilities • Trade off imbalance
4	TELLER A Trustworthy Framework For Explainable Generalizable and Controllable Fake News Detection	2024, arXiv	• Dual system framework • Neural symbolic models • Logical rule aggregation	• Data bias • System constraints • Manual refinement needed

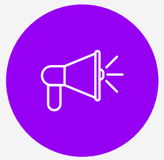
LITERATURE **SURVEY**

S.NO	TITLE	YEAR & PUBLICATION	METHODOLOGY USED	LIMITATIONS
5	EMIF Evidence aware Multi source Information Fusion Network for Explainable Fake News Detection	2024, arXiv	• Multi source fusion • Co attention network • Inconsistency loss	• Noise filtering • Model robustness • Cognitive bias
6	Explainable Artificial Intelligence Applications in Cyber Security State of the Art in Research	2022, IEEE Access	• XAI techniques • AI based cybersecurity • Rule extraction	• Benchmarking gaps • Cyber threats • Ethical concerns

LITERATURE **SURVEY**

S.NO	TITLE	YEAR & PUBLICATION	METHODOLOGY USED	LIMITATIONS
7	Advancing Fake News Detection Hybrid Deep Learning with FastText and Explainable AI	2024, arXiv	• Hybrid deep learning • FastText embedding • Post hoc explainability	• Dataset bias • Feature dependency • High complexity
8	When Explainability Turns into a Threat – Using XAI to Fool a Fake News Detection Method	2024, Elsevier – Computers and Security	• SHAP explainability • Adversarial attack • NLP model manipulation	• Security risks • Model deception • Explainability misuse

EXISTING **SYSTEM**

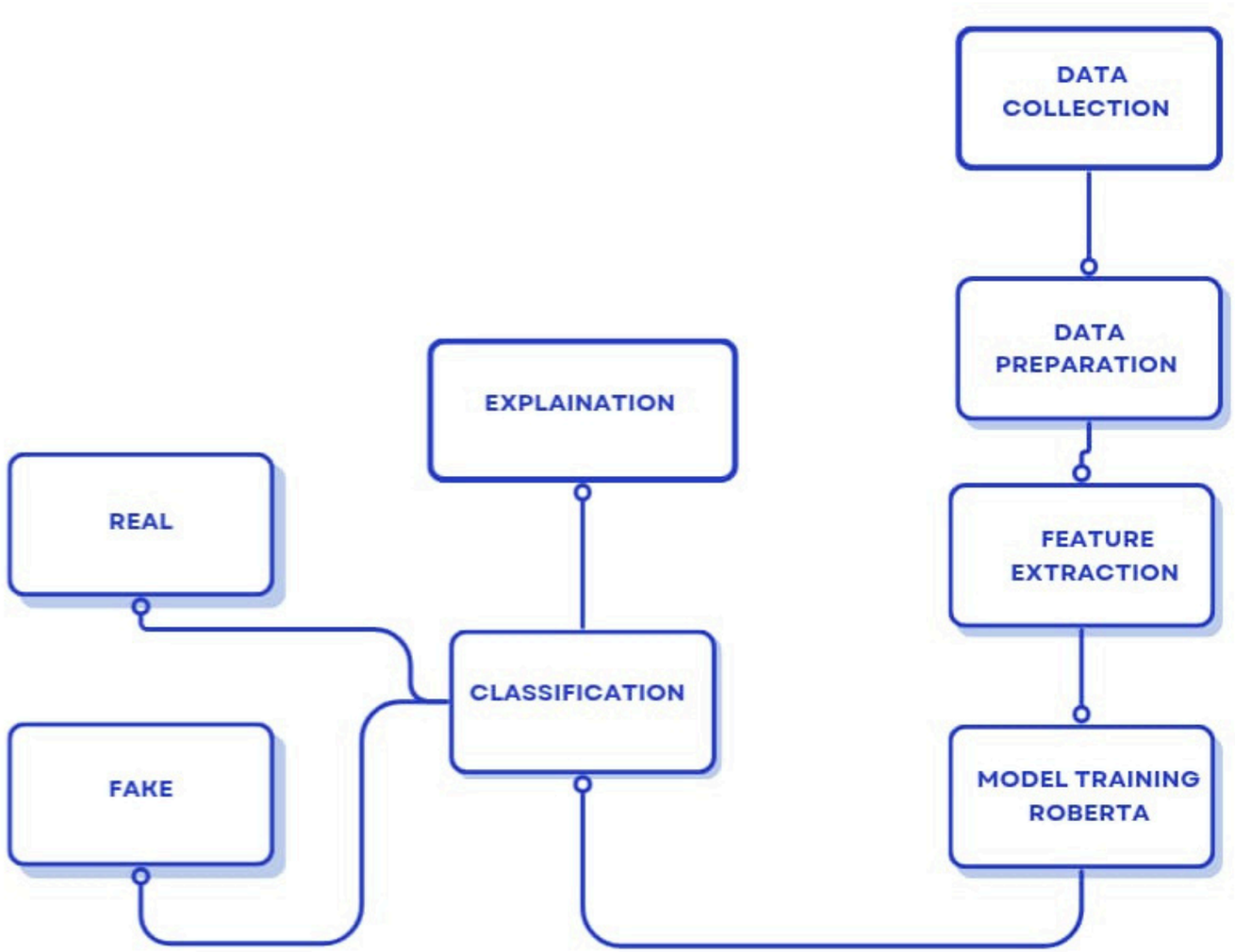


Rule-based approaches are ineffective against evolving fake news patterns

Machine learning models lack transparency, making decisions hard to interpret

Traditional NLP models struggle with contextual understanding.

BLOCK **DIAGRAM**



PROPOSED **SYSTEM**

Implementing RoBERTa for fake news classification

Using RoBERTa embeddings for feature representation

Applying LIME/SHAP for explainability.

MODULES

DATA PREPROCESSING

- **Bias-Aware Data Curation:** Collect data from diverse sources across political, cultural, and demographic lines to reduce representational bias and ensure inclusivity.
- **Sensitive Attribute Tagging:** Annotate data with labels for sensitive attributes (e.g., race, gender, region) to monitor potential biases and enable fairness-aware model training or auditing.
- **Text Cleaning & Normalization:** Apply ethical preprocessing steps like removing hate speech, offensive content, or misinformation artifacts while preserving context for accurate classification and fairness evaluation.

MODULES

DATA PREPARATION

- Removing stop words (e.g., "the," "is," "and") to reduce noise in the data.
- **Tokenization:** Splitting text into individual words or phrases.
- **Lowercasing:** Converting all text to lowercase for uniformity.
- **Removing special characters and numbers:** Keeping only meaningful words.
- **Stemming/Lemmatization:** Reducing words to their root form (e.g., "running" → "run").
- **Handling missing data:** Removing or filling in missing values.

MODULES

FEATURE EXTRACTION

- **Sentence Embeddings (RoBERTa):** Generates context-aware vector representations of entire sentences rather than individual words.

- **Text Encoding:** Input text is tokenized and passed through RoBERTa to generate rich, contextualized embeddings using its deep transformer layers.

- **Feature Representation:** The [CLS] token or averaged hidden states are extracted as fixed-length feature vectors for downstream tasks like classification or clustering

MODULES

MODEL TRAINING

- **Using more training data:** RoBERTa removes the Next Sentence Prediction (NSP) task and trains on more diverse datasets.
- **Applying dynamic masking:** Unlike BERT, which uses static masks, RoBERTa applies random masking to improve robustness.
- **Using larger batch sizes:** This leads to better generalization for NLP tasks.

MODULES

CLASSIFICATION

- Once trained, the model predicts whether a given news article is FAKE or REAL based on the learned representations.
- If the text contains misleading or false information, the model labels it as FAKE.
- If the text is trustworthy and factually accurate, the model labels it as REAL.

MODULES

EXPLANATIONS

- LIME and SHAP enhance model explainability.
- LIME perturbs input data to observe prediction changes, identifying influential features, while SHAP uses Shapley values to quantify each feature’s contribution to a prediction.
- Both methods provide transparent, human-understandable explanations, helping identify biases and ensuring fairness in deep learning models, making them more ethical and interpretable for decision-making and auditing.

CHALLENGES FACED **AND** SOLUTIONS

No proper benchmark datasets	Combined multiple datasets
Bias	Identification and Mitigation
Lack of completeness	Balanced and Stratified

MODELS **COMPARISON**

	Precison	Recall	Accuracy	F1 Score
LSTM	95.90	95.00	95.00	95.30
BERT	97.00	98.00	97.46	97.00
RoBERTa	98.82	98.56	98.34	98.82

USER **INTERFACE**

HTML, CSS

JavaScript

Flask

OUTPUTS

```
explainer = LimeTextExplainer(class_names=class_names)
exp = explainer.explain_instance(text, predictor, num_features=10,num_samples=1322)
exp.show_in_notebook(text=text)
```

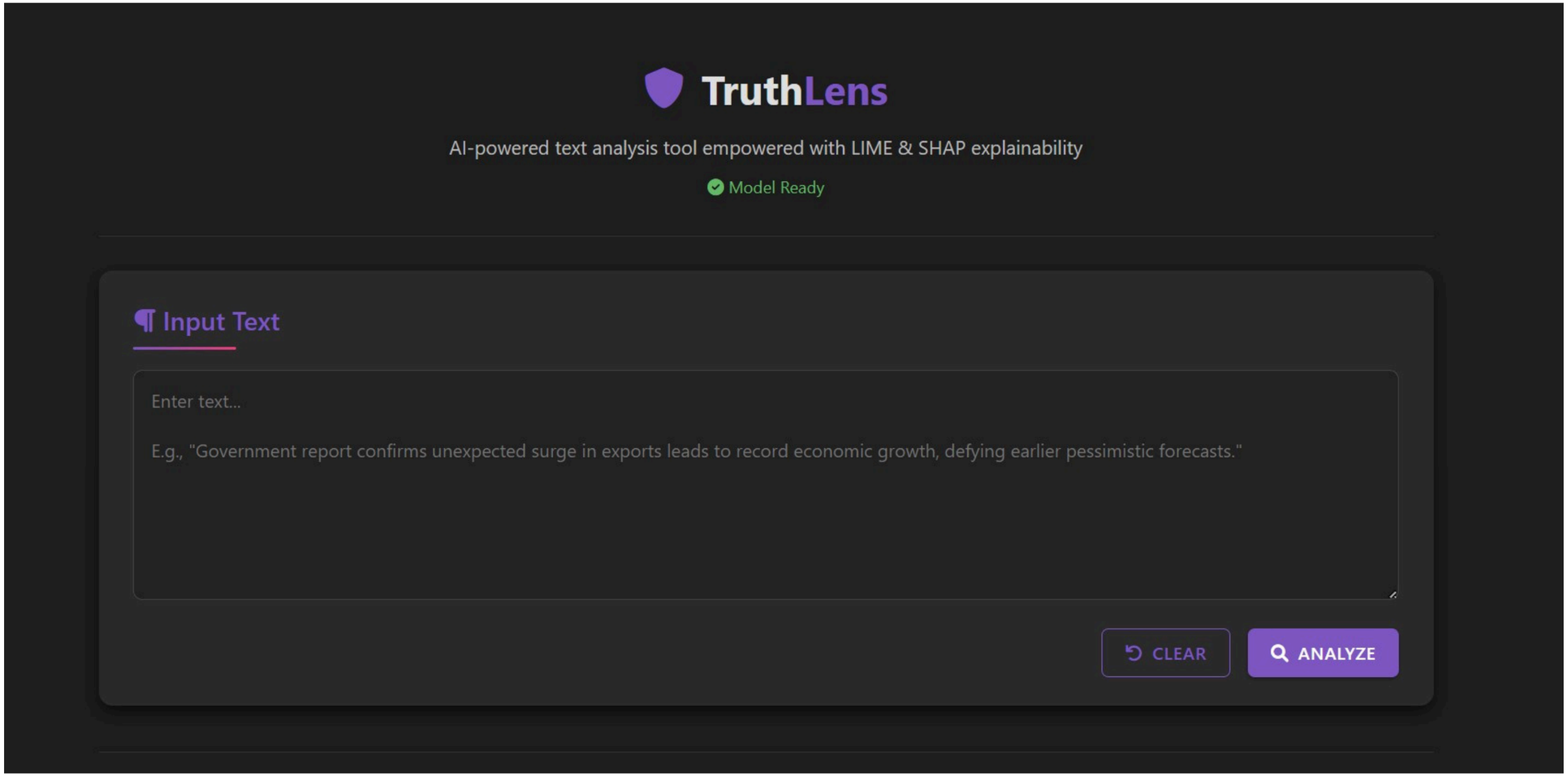

```
text = "A rendering of the Trump Tower in Makati City, Philippines MAKATI CITY, Philippines ( The Adobo Chronicles) – When completed by the end of 2016, the Trump T
exp = explainer.explain_instance(text, predictor, num_features=10,num_samples=2000, top_labels=1)
exp.show_in_notebook(text=text)
```

OUTPUTS


PartitionExplainer explainer: 2it [00:14, 14.39s/it]
outputs
Output 0 Output 1
f_output 0 (inputs)
0.473498
base value
0.48092
0.49
0.5
0.51
0.52
inputs
Er nie GEl s , Gknown Gas Gthe GBig GEasy , Gcreated Ga Glittle Gstrife Gon Gsocial Gmedia Gthis Gmonth Gby Gplaying Ggolf Gwith GPresident GTrump Gand GJapan âĢ Ļ s Gprime Gminister , GShinzo GAbe . Gat GTrump GNational GGolf GClub Gin GJupiter . GFla . GAs Gthe GP GA GTour Gmoves Gthis Gweek Gto GP GA GNational GResort Gand GSpa . Ga G G Gdrive Gfrom GTrump GNational , Git Gwas Gnatural Gto Gwonder : GHow Gmany Got GEl s âĢ Ļ s Gpeers Gon Gthe Gcour , Gif Gextended Gthe Gsame Ginvitation , Gwould Gsay Gyes ? GAt GRiv lera GCountry GClub Gin GLos GAngeles Glast Gweek . Gl Gasked Gmore Gthan G G Got Gthe GGenesis GOpen Gield Gthat Gquestion . Ggranting Gthe Grespondents Ganonymity Geo Gthey Gwouldn âĢ Ļ t Grisk Gthe Gwrath Got Gtheir GTwitter Gtollowers Gor . Gin Gthe Gcase Got Gat Gleast Gone Gpro . Gthe Gfury Gof Gnis Gwitle . GThe Gplayers Grange Gin Gage Gfrom Gthe Gearly G20 s Gto Glate G40 s . GThey Grepresent Gnine Gcountries Gand Gmake Gtheir Ghomes Gin G14 GAmerican Gstates , Gincluding Gfour Gthat Gvoted Goverwhelmingly GDemocratic Gin Glast Gyear âĢ Ļ s Gpresidential Gelection . GOf Gthe G56 Gplayers Gpolled . G50 G GâĢĶ G G Gor G89 . 3 Gpercent G GâĢĶ G G Gsaid Gthey Gwould Gplay Ggolf Gwith GTrump Gif Gasked . GOnly Gthree Gsaid Gthey Gwould Gnot . GThe Gremaining Gthree Gdeclined Gto Ganswer . GThe Gresults Gwere Ghardly Gsurprising . GThe Golub houses Gat GP GA GTour Gstops Ghave Glong Gtrend ed GRepublican , Gand Gthe Gsport âĢ Ļ s Gtarget Gdemographic G GâĢĶ G G Grich . Gmostly Gwhite Gmen G GâĢĶ G G Gis Gfar Gdifferent Gfrom Gthe Gwomen , Gminorities , Gimmigrants Gand GMuslims Gwho Ghave Gat Gtimes Gbeen Gthe Gmost Goffended Gby Gthe Gpresident âĢ Ļ s Gstatements Gand Gpositions .
Fake: 0.53
Real: 0.47
Final Prediction: Fake


PartitionExplainer explainer: 2it [00:10, 10.40s/it]
outputs
Output 0 Output 1
0.454057
0.48
0.48
0.49
0.5
0.51
0.52
0.53
0.54
inputs
Hillary GClinton GAppears GDis otented GAnd GConf used GAtGNew GYork GAirport GDay GBefore GPresidential GElection GA Gew Gseconds Gater , Gher Ghandlers Gbecome Gaware Gthat Gshe 's Gbeing Ga GParkinson 's Gdisea Gmotment . Gand Gimmediately Gswerv Garound Gher Gpushing Gthe Gvideo Gcameras Gbackwards . GBut Gnot Gbefore Gvideo Got Gher Gis Grecorded Gshooting Gshe Gdoesn 't Gknow Gwhat Gshe 's Gsupposed Gto Gdo .G G GAfter GHillary GClinton Gposed Gfor Gphotos Gwith Gthe Gmedia Gsoverning Gher , Gshe Gwas Gconfused Gabout Gwhether Gshe Gneeded Gto Gboard Gthe Gawaiting Golane Gor Ggot Gin GtheGmotor cade Gthat Ghad Gjust Gdropped Gher Goff . G C Hillary GClinton Gwas Gat Gthe GWhite GPlains , GNew GYork Gairport Gabout Gto Gstart Gher Gday . Gwhen Gshe Gexhibited Gsome Gstarting Gbehavior G Git Gwas Gthe Gvery Gbeginning Got Gher Gday , Geo Gpresumably Gshe Gwas Created Gand Gdirt Gas Gshe Gprepared Gto Grit Gthe Gcampaign Gtrail
Fake: 0.46
Real: 0.54
Final Prediction: Real

OUTPUTS


TruthLens
AI-powered text analysis tool empowered with LIME & SHAP explainability
Model Ready
Input Text
Enter text...
E.g., "Government report confirms unexpected surge in exports leads to record economic growth, defying earlier pessimistic forecasts."
CLEAR
ANALYZE

OUTPUTS

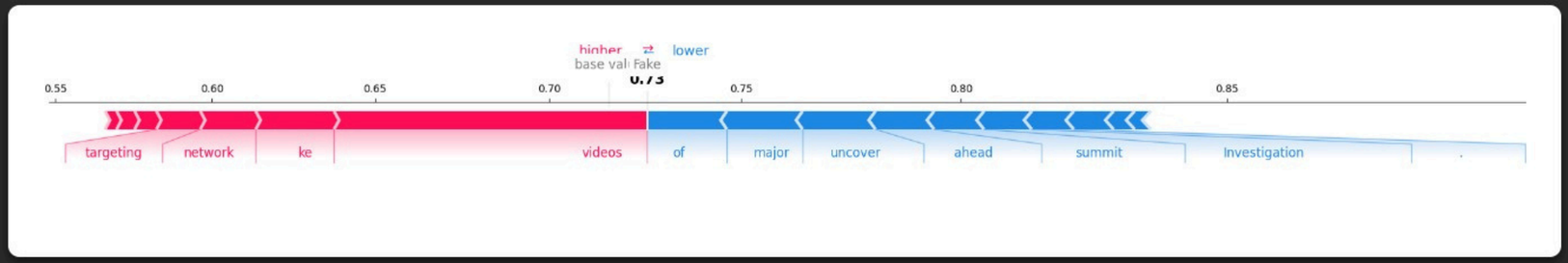
LIME Explanation

HIGHLIGHTED TEXT | KEY FEATURES | VISUAL PLOT

Investigation uncovers network of deepfake videos targeting political figures ahead of major summit .

SHAP Explanation

This plot shows how each word/token pushes the model's prediction. Red pushes towards 'FAKE', Blue pushes towards 'REAL'. The length of the bar indicates the magnitude of influence.



PAPER **ACCEPTANCE**

--------- Forwarded message ---------
From: **ISETE** <info.iseteconference@gmail.com>
Date: Tue, 6 May, 2025, 4:26 pm
Subject: PAPER ACCEPTANCE CONFIRMATION
To: aravind.2111006@srec.ac.in <aravind.2111006@srec.ac.in>

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.

CONCLUSION

- The RoBERTa-based fake news detection model demonstrates high accuracy and reliability in identifying misinformation by leveraging contextual language understanding.
- It outperforms traditional models in distinguishing between real and fake news.
- The approach is both efficient and scalable for real-time applications.
- Ethical considerations and bias mitigation enhance its trustworthiness.

REFERENCES

[1]. C.-H. Chuan et al., “Explainable Artificial Intelligence (XAI) for Facilitating Recognition of Algorithmic Bias,” Telematics and Informatics, vol. 91, 2024.

[2]. R. Kozik et al., “When Explainability Turns Into a Threat Using xAI to Fool a Fake News Detection Method,” Computers & Security, vol. 137, 2024.

[3]. H. Liu et al., “TELLER: A Trustworthy Framework for Explainable, Generalizable and Controllable Fake News Detection,” arXiv preprint, arXiv:2402.07776v2, 2024.

[4]. [B. Wang et al., “Explainable Fake News Detection With Large Language Model via Defense Among Competing Wisdom,” in Proc. ACM Web Conf. (WWW), 2024.

REFERENCES

[5]. Q. Dong et al., “EMIF: Evidence-aware Multi-source Information Fusion Network for Explainable Fake News Detection,” arXiv preprint, arXiv:2407.01213v1, 2024

[6]. E. Hashmi et al., “Advancing Fake News Detection: Hybrid Deep Learning with FastText and Explainable AI,” IEEE Access, vol. 12, Mar. 2024.

[7]. J. Kathiriya and S. Degadwala, “A Review on Fake News Detection using Deep Learning Methods,” Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol., vol. 10, no. 3, 2024.

[8]. Z. Zhang et al., “Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research,” IEEE Access, vol. 10, 2022.

FUTURE **SCOPE**

- Future work may involve integrating multimodal data like images and videos to improve detection accuracy.
- Domain adaptation techniques can make the model effective across different topics and languages.
- Enhancing explainability will increase user trust and model transparency.
- Continuous bias auditing is essential for ethical AI deployment.

PLAGERISM **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.

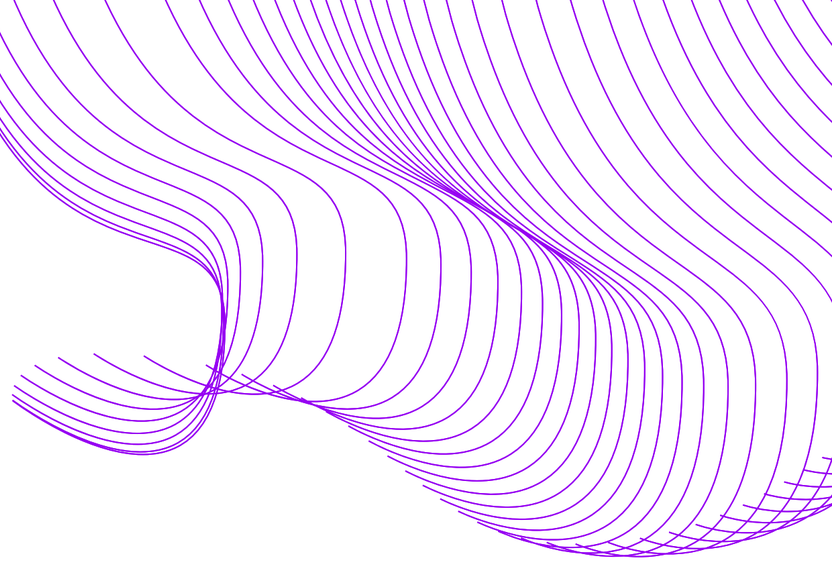
False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

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.



THANK **YOU**