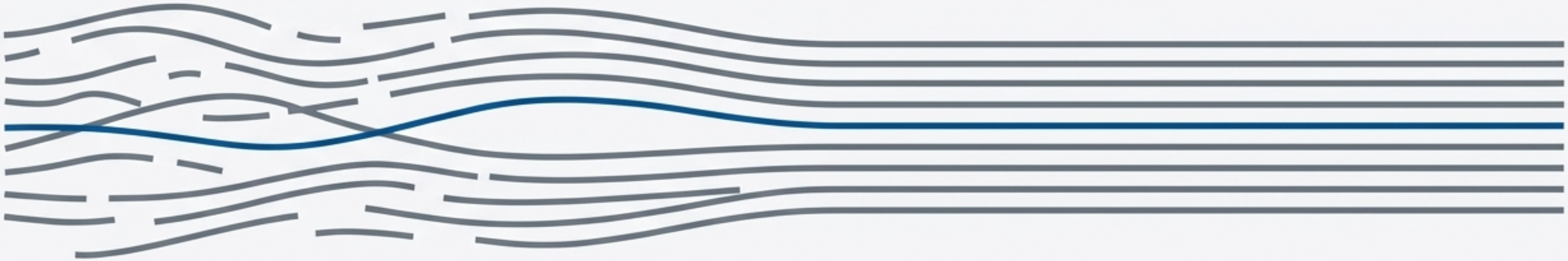
IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Fine-Tuned Transformers Achieve 93% F1-Score in Social Bot Detection

A deep dive into using transfer learning and explainable AI to identify sophisticated automated accounts on social media.



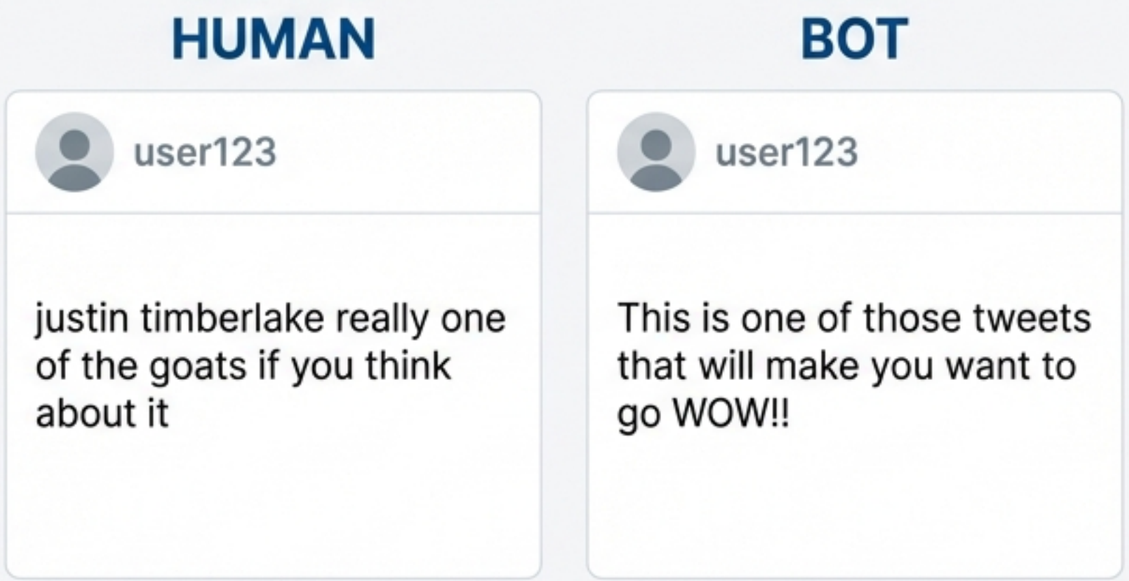
- Transfer Learning
- Pre-trained Language Models (PLMs)
- BERT & GPT
- Explainable AI (XAI)

Modern Social Bots Can Be Indistinguishable from Human Users

"It is becoming more and more difficult to discern a text produced by deep neural networks from that created by humans."

- Advanced bots powered by models like GPT-3 generate fluent, coherent, and grammatically accurate text.
- This sophistication allows them to infiltrate online communities, disseminate misinformation, and manipulate discourse in politics, finance, and public health.
- The short-form nature of platforms like Twitter (280 characters) makes detection even harder, as brief messages offer fewer linguistic cues.

Even Twitter's internal studies estimate that up to 5% of its active users are fraudulent or spam bots.



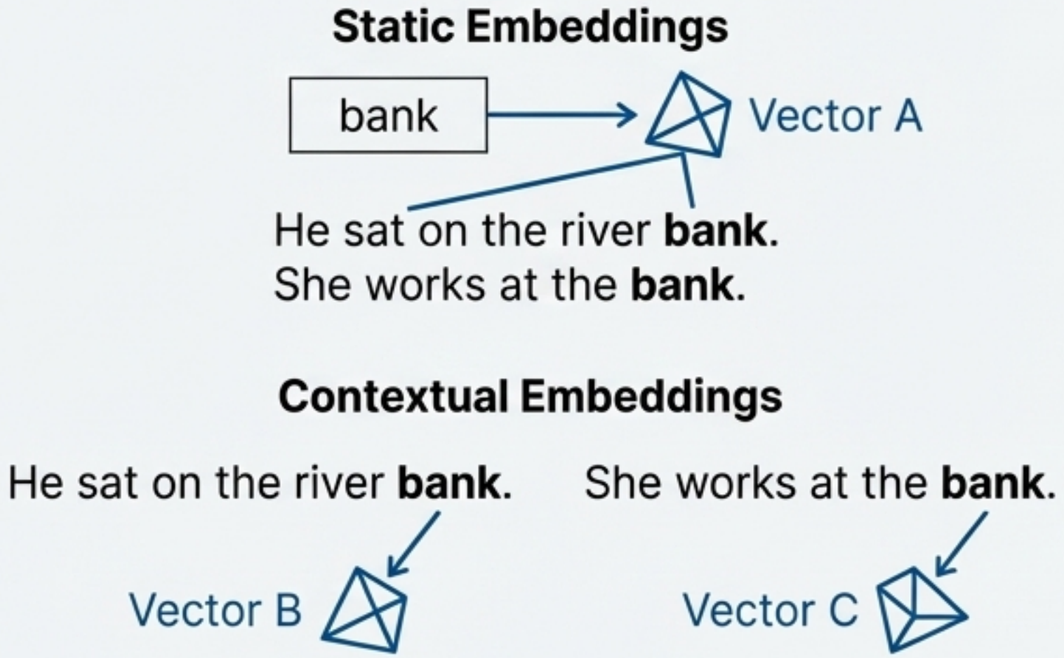
Traditional Detection Methods Struggle to Keep Pace with Evolving Threats

The Challenge: Bot detection methods are often not robust enough to generalize beyond their specific training data.

Limitations of Previous Approaches

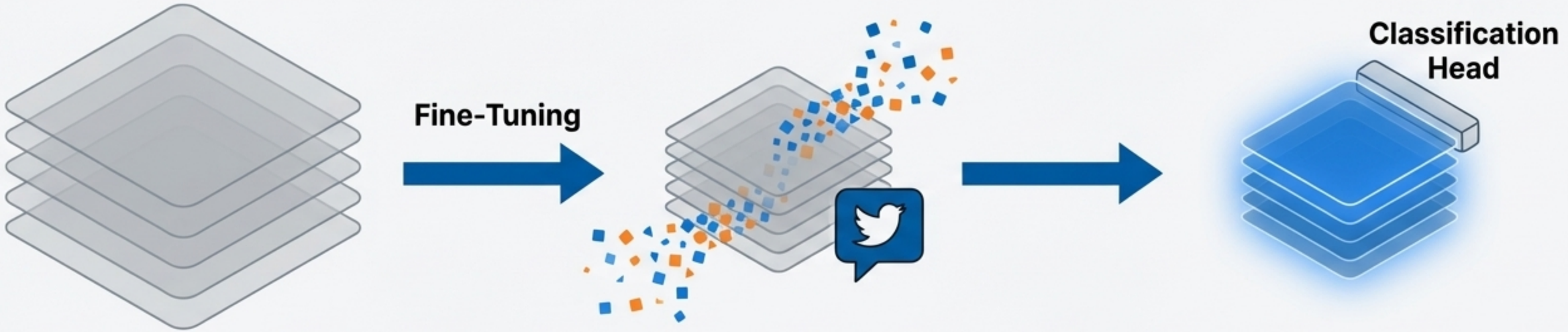
- **Static Word Embeddings (Word2Vec, GloVe):** Generate a single, static vector for each word, failing to capture context.
- **Reliance on Feature Engineering:** Traditional models require laborious, manual feature engineering, which is slow and may miss subtle linguistic patterns.
- **Lack of Domain-Specific Training:** Models are often pre-trained on generic text and struggle with the unique jargon and structure of social media posts.

The Contextual Shift



Our Approach: Fine-Tuning Pre-trained Language Models on Bot-Specific Data

Core Concept: We leverage the power of large, pre-trained language models (PLMs) and adapt their deep understanding of language to the specific task of bot detection.



Powerful PLM (e.g., BERT, GPT-3) Pre-trained on a massive, general text corpus.

The model is fine-tuned on a specialized dataset of bot and human tweets, adjusting its weights to learn the unique linguistic patterns of bot-generated content.

A classification head makes the final 'Human" or 'Bot" prediction based on the specialized model's output.

Benefit: This method obviates the need for manual feature engineering, as the model learns the relevant features automatically.

We Evaluated a Suite of State-of-the-Art Transformer Architectures

Transformers excel at understanding language context through mechanisms like self-attention, making them ideal for this task. We compared several leading models.

BERT-Based Models

BERT (base & large)

The original bidirectional model that revolutionized NLP. Learns from both left and right context.

RoBERTa

A robustly optimized BERT with improved training methodology and larger datasets.

DistilBERT

A smaller, faster, and more efficient version of BERT that retains over 95% of its performance.

XLM-RoBERTa

A multilingual model pre-trained on 100 languages.

GPT-Based Models

GPT-3

A powerful generative model known for its text embedding capabilities, which create high-quality vector representations of text.

The Proposed Model Follows a Four-Step Detection Pipeline

Step 1: Data Acquisition & Preprocessing

Datasets: TweepFake, fox8-23.

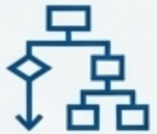
Cleaning: Remove noise, standardize URLs/mentions with tags (<url>, <user>), tokenize text.

Step 2: Fine-Tuning PLMs

Input: Cleaned tweet text.

Process: Fine-tune selected transformer models (BERT, RoBERTa, etc.) on the training dataset (80% split).

Output: A specialized model that generates contextualized embeddings.



Step 3: Classification

Input: The contextualized embedding vector from the fine-tuned PLM.

Process: Pass the vector through a two-layer Feedforward Neural Network (FNN).

Step 4: Final Prediction

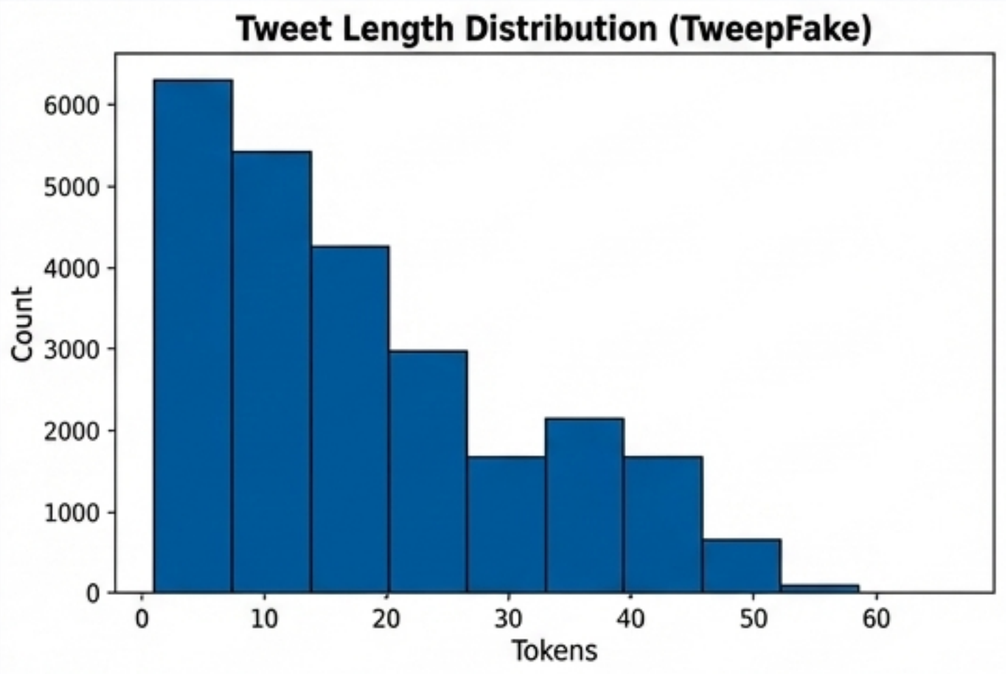
Output: A binary classification of 'Human' or 'Bot'.

The Models Were Fine-Tuned and Validated on Two Public Datasets

TweepFake

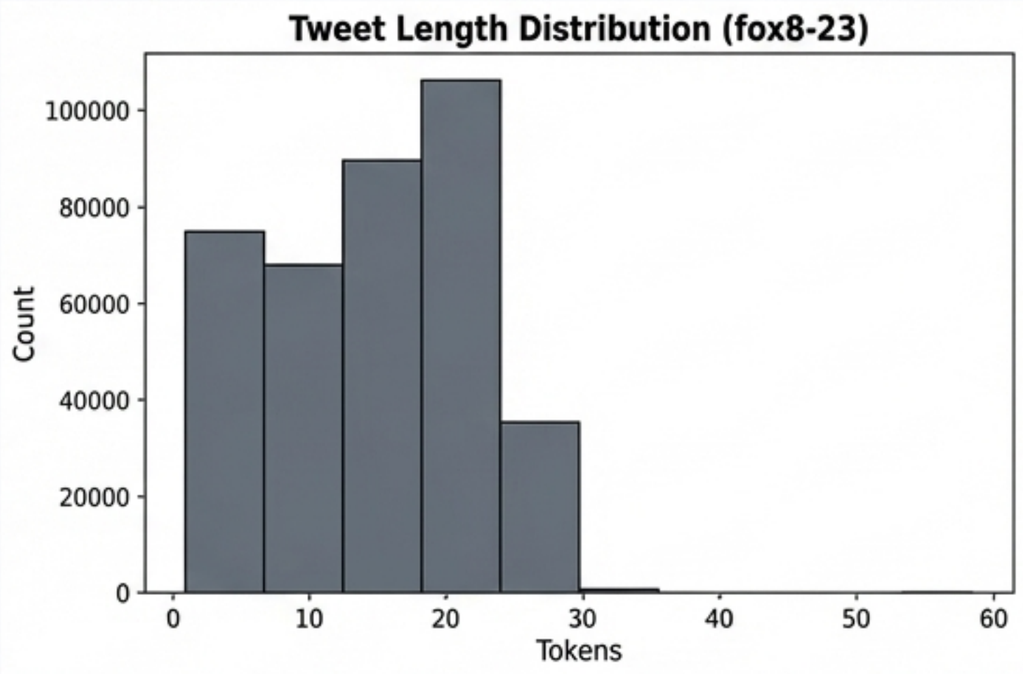
Composition: 25,572 tweets, perfectly balanced between human-generated and bot-generated content.

Bot Content Source: Generated by various techniques including Markov Chains, RNN, LSTM, and GPT-2.



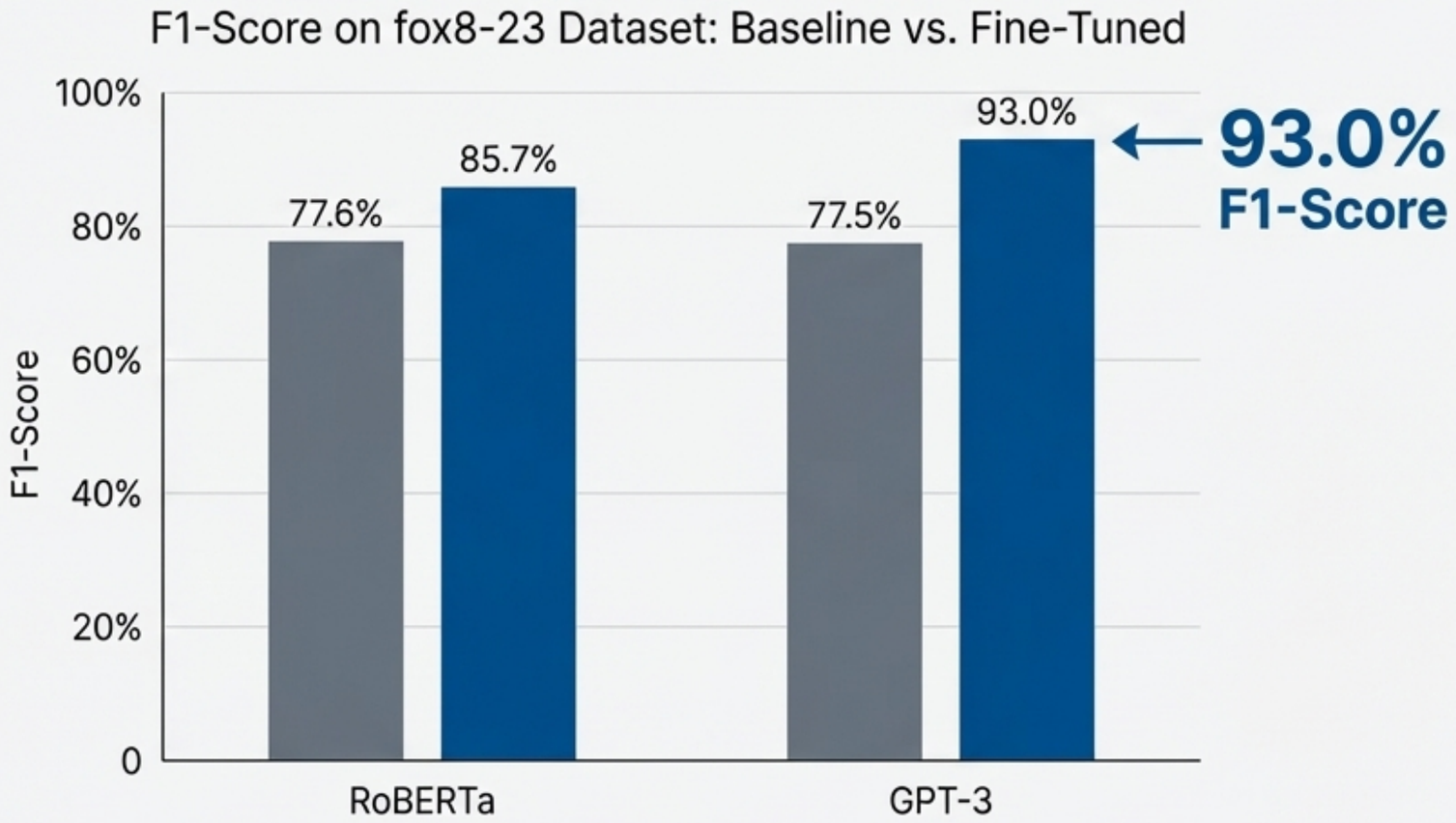
fox8-23

Composition: 218,245 human tweets and 149,783 bot tweets from 1,140 human and 1,140 bot accounts.



Data Split: Datasets were split 80% for training and 20% for testing, with no overlapping users between sets to ensure fair evaluation.

Fine-Tuned Models Dramatically Outperform Baselines, Achieving Up to 93% F1-Score



Peak Performance: The fine-tuned GPT-3 model achieved an F1-score of 92.98% and accuracy of 93.07% on the fox8-23 dataset.

Consistent Improvement: Fine-tuning improved the F1-score of RoBERTa by 14% and GPT-3 by nearly 20%.

State-of-the-Art: This performance surpasses previous benchmarks reported for both datasets.

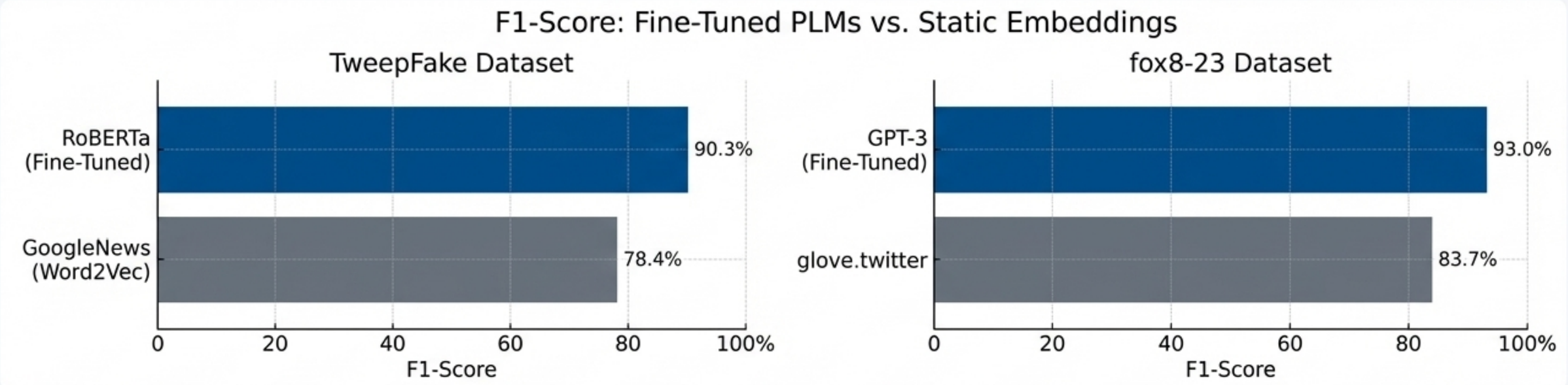
Performance Varies Across Models, but Fine-Tuning Consistently Boosts Results

Model	Dataset	Fine-Tuned Model Performance	AUC	F1-Score
BERT (base)	TweepFake	0.8989	0.9365	0.8907
BERT (large)	**TweepFake**	**0.9407**	0.9243	0.8887
RoBERTa	**TweepFake**	0.9053	**0.9660**	**0.9029**
XLM-RoBERTa	TweepFake	0.8823	0.9642	0.8821
DistilBERT	TweepFake	0.8750	0.8743	0.8749
GPT-3	TweepFake	0.8762	0.9307	0.8809
BERT (base)	fox8-23	0.9150	0.9703	0.9075
BERT (large)	fox8-23	0.9264	0.9083	0.8614
RoBERTa	fox8-23	0.8566	0.8873	0.8566
XLM-RoBERTa	fox8-23	0.8087	0.8683	0.8127
DistilBERT	fox8-23	0.9284	**0.9768**	0.9274
GPT-3	fox8-23	**0.9307**	0.9740	**0.9298**

- On the **TweepFake** dataset, **RoBERTa** achieved the highest F1-score (90.29%) and AUC (96.6%).
- On the more complex **fox8-23** dataset, **GPT-3** was the top performer across Accuracy and F1-Score.
- The results show that while baseline models are proficient, fine-tuning unlocks a significant, measurable improvement by adapting them to domain-specific intricacies.

Fine-Tuned PLMs Are Superior to Traditional Static Embeddings

The Experiment: We compared our fine-tuned models to a BiLSTM classifier using pre-trained static embeddings (GloVe and Word2Vec), a common traditional approach.



Key Finding: Dynamic, contextual embeddings from transformer-based PLMs provide a richer representation of text, leading to better classification performance.

Explanation: *"Unlike Word2Vec and GloVe, which generate a single embedding vector for each word... transformer-based PLMs provide dynamic embeddings. This means that the representation of a word changes based on its context within a sentence."*

We Can Look Inside the “Black Box” to Understand Model Decisions

The Challenge:

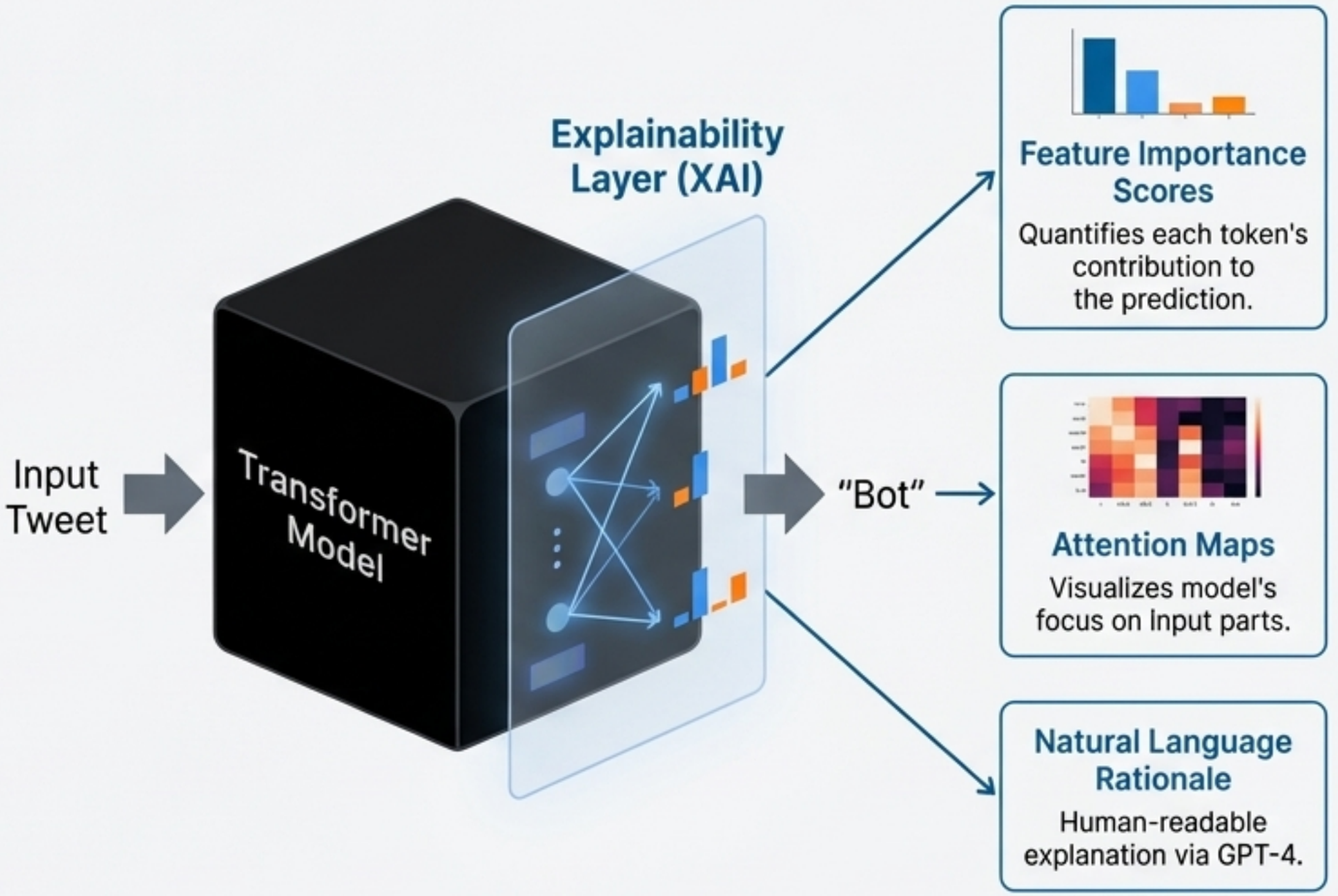
Transformer models are highly effective but their complexity can make it difficult to understand *why* they make a specific prediction.

Our Solution:

We integrated Explainable AI (XAI) techniques to dissect and interpret the model's decision-making process, enhancing transparency and reliability.

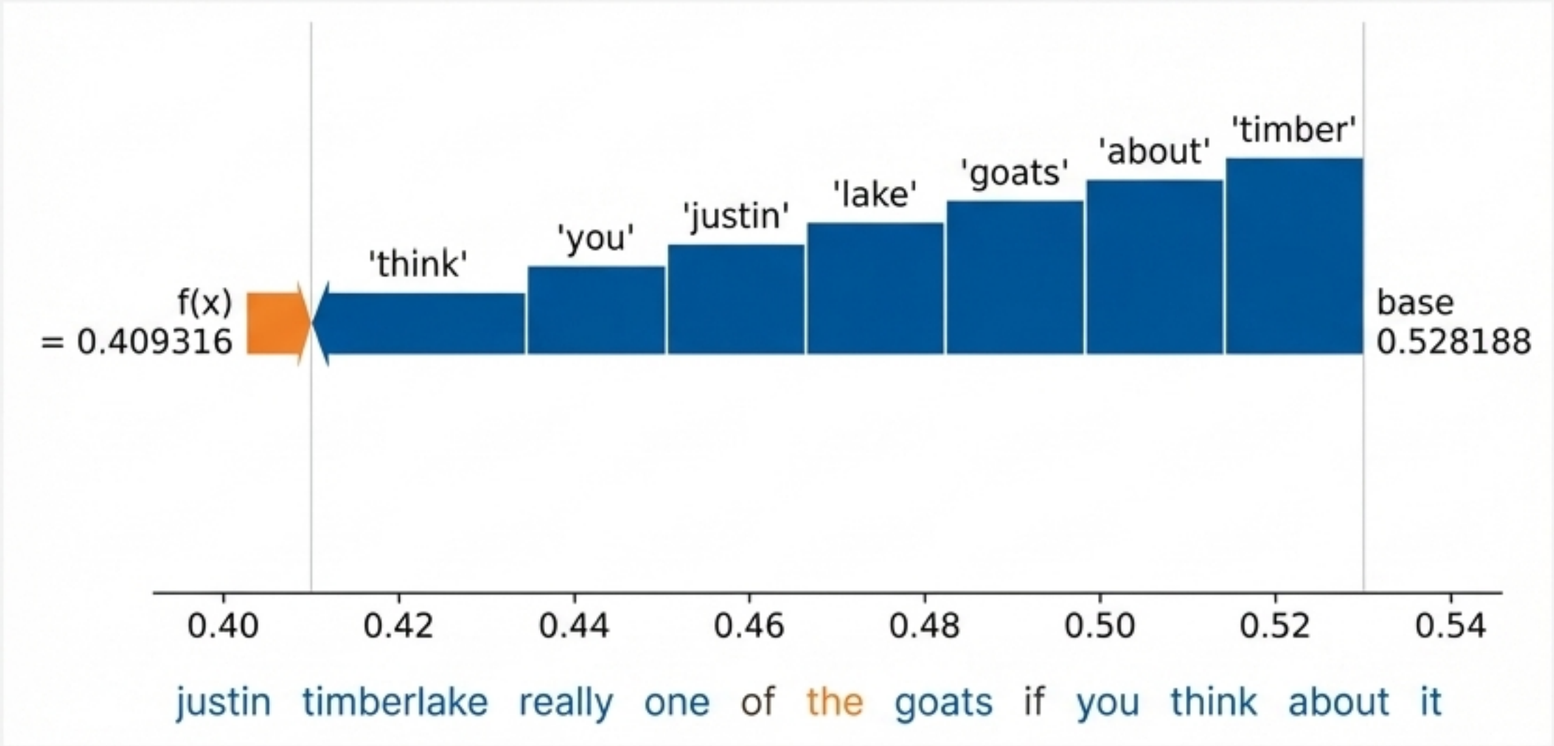
Techniques Used:

1. **SHAP (SHapley Additive exPlanations):** Quantifies the contribution of each word (token) to the final prediction.
2. **Attention Visualizers:** Shows which parts of the input text the model 'focuses on' when processing information.
3. **LLM-based Prompting:** Uses GPT-4 to generate human-readable, natural language explanations for a classification.



SHAP Analysis Reveals the Specific Linguistic Cues Driving Predictions

Case Study: Classifying the tweet *"justin timberlake really one of the goats if you think about it"* as human-generated using the fine-tuned DistilBERT model.



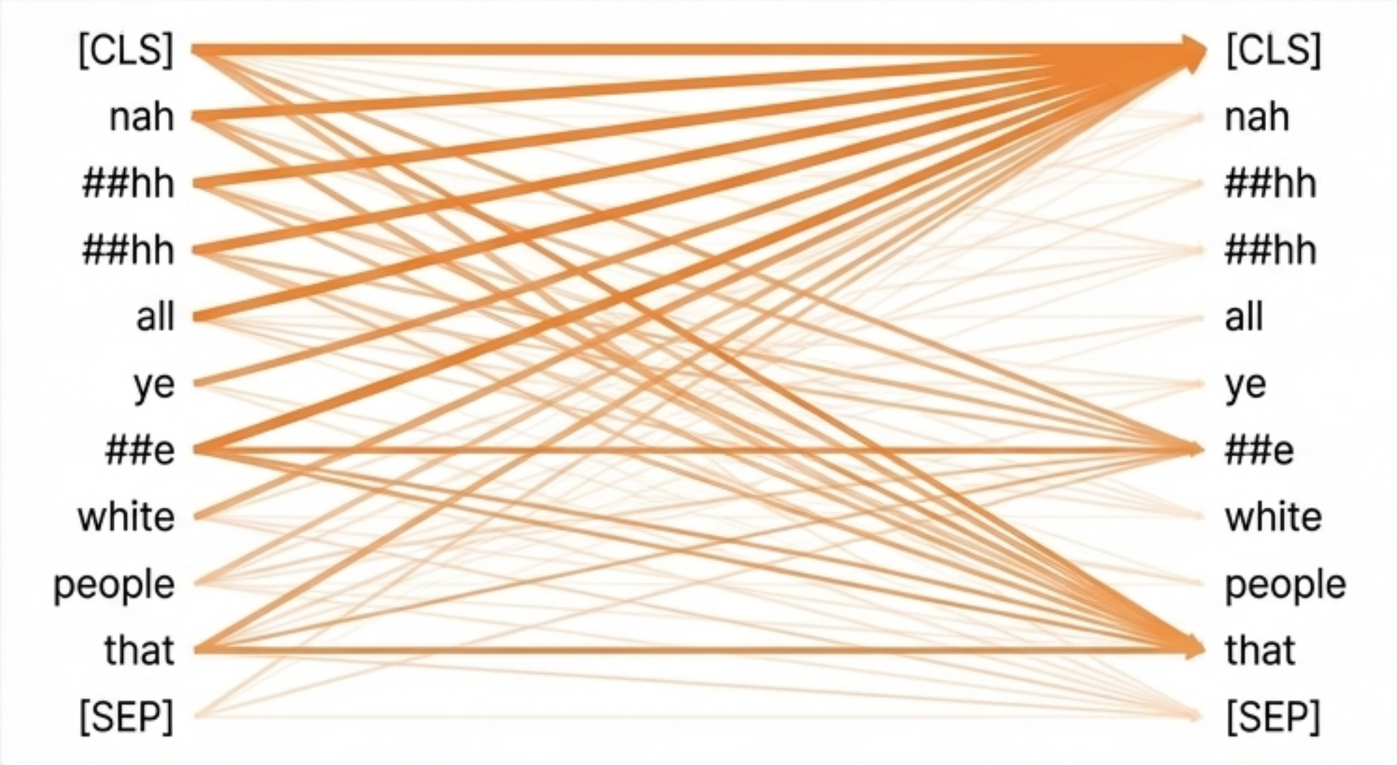
Interpretation:

- The plot shows the model's output value (0.409) is lower than the base value (0.528), pushing the prediction towards 'Human'.
- **Blue features (decrease prediction):** Words like 'justin', 'timberlake', and 'think' strongly indicate human origin. The model has learned that personal names and cognitive verbs are less common in bot-generated text.
- **Red features (increase prediction):** Common words like 'the' slightly push the prediction towards 'Bot', but are outweighed by the stronger human signals.

Takeaway: The fine-tuned model learns to discern nuanced linguistic cues that distinguish between human and bot-generated text.

Attention Visualizations Show What Parts of a Tweet the Model Focuses On

Case Study: Analyzing the attention patterns of the XLM-RoBERTa model on the bot-generated tweet *“nahhhhh all yee white people that”*.



Interpretation:

- This view from Layer 2, Head 2 shows a pronounced focus on the `[CLS]` token, which aggregates the overall meaning of the sentence for classification.
- Substantial attention is given to the token '##e', suggesting its significance in the tweet's context as learned by the model.
- The model processes subword elements like '##hh' and '##e', demonstrating its ability to handle unconventional language and spelling common in social media.

Takeaway: Attention patterns reveal how the model deconstructs text and assigns importance to different tokens, which is critical for understanding its reasoning.

LLM-Prompting Translates Model Logic into Natural Language Explanations

Concept: We use GPT-4 with a zero-shot prompting strategy to automatically generate textual feedback that explains a classification decision to a user.

Tweet Classified as BOT

yo this fall weather is poison and socially progressiveâ=C that sack, sis twitter activists, and i just me?

GPT-4 Generated Feedback

The tweet exhibits several characteristics that are common in bot-generated content, including **unusual phrasing and syntax errors** ("socially progressiveâ=C", "that sack, sis"). These irregularities suggest automated content generation... Additionally, the tweet seems to **randomly combine different topics**, which is a common trait in bot-generated content...

Benefit: This provides accessible, context-aware explanations that enhance the interpretability and trustworthiness of the detection system for end-users.

Fine-Tuning and Explainability Mark a New Frontier in Bot Detection

Summary of Contributions

1. **Superior Performance**
   Demonstrated that fine-tuning PLMs on bot-specific data achieves state-of-the-art results (93% F1-score), significantly outperforming traditional methods.

2. **Enhanced Transparency**
   Proved that XAI techniques like SHAP, attention visualization, and LLM-prompting can make complex transformer models interpretable and trustworthy.

3. **Advanced Representation**
   Showcased the power of dynamic, contextual embeddings from models like BERT and GPT-3 over static embeddings for nuanced tasks.

Future Work

- **Multilingual Support:** Refine embedding mechanisms to support languages other than English, expanding the model's global applicability.
- **Hybrid Approaches:** Integrate content-based methods with graph-based classification by incorporating rich datasets like TwiBot-20 and TwiBot-22 to model user relationships and network topology for more robust detection.