SUNWAY UNIVERSITY

A Capstone Project Viva Presentation

Bot or Human?
Detection of DeepFake text on Twitter with Semantic, Emoji, Sentiment and Linguistic Features

by

ALICIA CHONG TSUI YING (20074290)

Bachelor of Information Systems (Honours) (Data Analytics)

Supervisor: Assoc. Prof. Dr. Chua Hui Na

CONTENTS

CONTNETS

INTRODUCTION

Background, Problem Statement, Research Objectives
Research Questions, Scope of Study, Contributions

INTRODUCTION

BACKGROUND

What is Deepfake text?

- refers to text that is created using AI and deep learning algorithms, imitating human writing style to a remarkable degree.
- In this work, deepfake text is also known as **machine-generated text**, in short, we call it **MGT**.
- Example: text generated by ChatGPT (AI chatbot), GPT-2, GPT-3, GPT-4 (language models)

BACKGROUND

Deepfake text: Applications

- has brought about both **promising advancements** and **potential risks** in the digital landscape.
- **legitimate applications:**
  e.g. creative writing, text summarization, information processing
- **misuse**:
  e.g. **spread of misinformation** and **fake news**
  - can have severe consequences on society
    - especially within the context of **social media** platforms

BACKGROUND

Twitter and Deepfake text

- Social media platforms, such as **Twitter**,
  - connecting millions of users worldwide and **facilitating the rapid exchange of information and ideas.**
- However, the proliferation of **deepfake text** poses a significant challenge to the **authenticity and reliability of the content shared** on these platforms.
- With the aid of social media **bots**, **deepfake text** can be potentially **shared on a large scale** to manipulate the public's opinion.

PROBLEM STATEMENT

Existing Deepfake text detection methods & tools

- manual **human evaluation and labeling**
  - impractical and prone to error
- Popular detectors: **GPTZero** & **Open AI's detector**
  - **are inadequate to handle the detection of short social media texts**, which are prevalent on platforms like Twitter
  - they require a minimum text length of 250 characters

PROBLEM STATEMENT

As a result,

- there is **a need for innovative research** to develop new effective methods that target the **detection of deepfake text on Twitter.**
- methods must be designed to **address the unique constraints and characteristics of the platform**, such as the **limited text length.**
- **Exploring novel features** from linguistic patterns, and contextual cues specific to social media text **holds promise in developing detection techniques for this task.**

PROBLEM STATEMENT

- **Building upon previous works** (Gambini et al., 2020; Saravani et al., 2021; Tesfagergish et al., 2021), **our research is centered around detecting short deepfake text samples on Twitter using the TweepFake dataset** introduced by Fagni et al. (2021).
- **While previous studies have primarily focused on tweet semantic text content, our aim is to develop a more robust detector by incorporating additional features**.
- These **additional features** include **emoji, linguistic and sentiment features** derived from the tweet content
  - drawing inspiration from works by Dukic et al. (2020), Dickerson et al. (2014), Fröhling & Zubiaga (2021), Hamida et al. (2022), and Heidari & Jones (2020).

PROBLEM STATEMENT

INTRODUCTION

- **To find relevant features for our detection model**, we aim to conduct **exploratory analysis** of **linguistic** and **sentiment** features to find key **differences between machine-generated text (MGT) and human-written text (HWT).**
- There is **limited research that examines the distinctions between traditional and modern machine-generated text** on social media
  - specifically text no longer than 280 characters.
- Therefore, we also aim to explore inherent differences between **traditional machine-generated text (TMGT)** and **modern machine-generated text (MMGT).**

TERMINOLOGY

- **Machine-Generated Text (MGT)**
  - deepfake text, created using artificial intelligence algorithms and deep learning models
- **Human-Written Text (HWT)**
  - text written by humans without the assistance of AI algorithms
- **Traditional Machine-Generated Text (TMGT)**
  - text generated by **simple** text-generative models
    - based on RNN, LSTM and Markov Chains architectures
- **Modern machine-generated text (MMGT)**
  - text generated by **advanced** text-generative models
    - based on the **Transformer** architecture.

RESEARCH QUESTIONS (RQS)

RQ1) What are the distinguishing characteristics between machine-generated text (**MGT**) and human-written text (**HWT**) on Twitter?

RQ2) What are the distinguishing characteristics between modern machine-generated text (**MMGT**) and traditional machine-generated text (**TMGT**) on Twitter?

RQ3) To what extent does incorporating **linguistic features, sentiment features**, and **emojis embeddings** alongside **semantic word embeddings** enhance the model's ability to accurately classify **MGT** and **HWT** on Twitter?

CONTRIBUTIONS

- **Advancement in Text Classification**:
    - Our findings highlight the **potential of leveraging semantic embeddings and supplementary features** including linguistic features and emoji features **to enhance the performance of deepfake text detector**.

- **Provided insights into linguistic and sentiment characteristics differences** between
    - **MGT** and **HWT**
    - **MMGT** and **TMGT**

- **Enhanced the TweepFake dataset** by Fagni et al., (2021)
    - by including additional MGT from the latest text generative models such as GPT-3.

LITERATURE
REVIEW

Related Works

LITERATURE REVIEW

#	Related Research	Best Approach Summary	Dataset	Embeddings & Features	Feature-based Approach	NLM Approach	Evaluated against	Performance
1	Fagni et al. (2021)	Fine-tuned RoBERTa model	TweepFake (Fagni et al., 2021)	Semantic Embeddings		✓	RNN, LSTM, Markov Chain, GPT-2	Acc: 89.6% $F1_{Bot}$: 89.7% $F1_{Human}$: 89.5%
2	Saravani et al. (2021)	BERT (word embeddings) + BILSTM (capture temporal relations) + NeXtVLAD (parametric pooling area)	TweepFake	Semantic Embeddings		✓	RNN, LSTM, Markov Chain, GPT-2	Acc: 92% $F1_{Bot}$: 92% $F1_{Human}$: 92%
3	Gambini et al. (2020)	Fine-tuned GPT-2-based classifier	TweepFake	Semantic Embeddings		✓	RNN, LSTM, Markov Chain, GPT-2	Acc: 91%
4	Tesfagergish et al. (2021)	Fine-tuned RoBERTa (word embeddings) + Hierarchical Attention Network (classifier)	TweepFake	Semantic Embeddings		✓	RNN, LSTM, Markov Chain, GPT-2	Acc: 89.7% F1: 85.5%
5	D. Dukic et al. (2020)	BERT Base model (word embeddings) + emoji2vec (emoji embeddings) + Logistic regression (classifier)	PAN dataset	Semantic Embeddings, Tweet metadata, Emoji embeddings	✓	✓	Unknown	F1: 83.36%
6	Heidari and Jones (2020)	Fine-tuned BERT (sentiment features) + GloVe (word embeddings) + neural network (classifier)	Cresci et al.'s dataset	Semantic Embeddings, Sentiment features	✓	✓	Unknown	Acc: 94% F1: 94.7%
7	Fröhling et al. (2021)	Ensemble of Classifiers (Logistic Regression, SVM, Random Forest, Neural Network)	Long text datasets	Linguistic features	✓	-	GPT-2, GPT-3, Grover	(Multiple results)
8	Hamida et al. (2022)	Deep Neural Autoencoder (linguistic features) + GloVE - BiLSTM autoencoder (word embeddings) + BiRNN (classifier)	Cresci et al.'s dataset	Semantic Embeddings, Linguistic features	✓	-	Unknown	Acc: 92.22% F1: 92%
9	Dickerson et al. (2014)	Ensemble of Classifiers (SVM, Gaussian naïve Bayes, AdaBoost, Gradient Boosting, Random Forest, Extremely Randomized Trees)	2014 India Election Dataset	Sentiment features, Tweet Syntax, User Behaviour	✓	-	Unknown	-
	Our proposed model	Fine-tuned BERT (word embeddings) + emoji2vec (emoji embeddings) + linguistic features + Neural Network (classifier)	**Enhanced TweepFake** *	Semantic Embeddings, Emoji embeddings, Sentiment features, Linguistic features	✓	✓	RNN, LSTM, Markov Chain, GPT-2, **GPT-3**	-

Note: TweepFake denotes the newly Enhanced TweepFake dataset used in this work. NLM denotes complex neural language models.*

METHODOLOGY

Dataset, Exploratory Data Analysis, Modeling Experiments

DATASET

TweepFake dataset

- a **Twitter deepfake text dataset** created by **Fagni et al. (2021).**

- comprise of human and machine-generated tweets.

- Unlike other datasets that rely on human annotations, this dataset was **compiled by manually selecting tweets** from **genuine human accounts** and **their corresponding fake bot counterparts**
  - this ensured the reliability of the data labels

DATASET

Original TweepFake dataset

- consists of **25,572 tweets**
- scraped from **23 human and bot accounts** on Twitter in 2021
- Text generative technologies used by bot accounts:
  - RNN, LSTM, Markov Chain and GPT-2.
- **Dataset** was **published** with only ***tweet_ID*** and ***label***.
  - Therefore, we are required to **scrape the tweets content**
- However, about **30% of the tweets were no longer accessible**
  - tweet deleted or account deactivated.
- To address this limitation, we created the **Enhanced TweepFake dataset.**

DATASET

Enhanced **TweepFake dataset**

- Dataset includes **supplementary tweet data** from **newly identified bot and human accounts** on Twitter
  - thereby **augmenting** the original **TweepFake dataset.**
- **Text generative technologies** used by **newly identified bot accounts: GPT-3, ChatGPT.**

Dataset	Subset of Dataset	Tweets	Tweets
Original TweepFake	-	-	25572
Enhanced TweepFake	Usable Tweets from Original TweepFake	17604	21730
	New Tweets to complement Original TweepFake	4126	

METHODOLOGY

DATASET

Enhanced TweepFake dataset

Account Type	Accounts	Tweets
Bot	24	10865
Human	17	10865
Total	41	21730

<table>
<tr><th>MGT Type</th><th>Technology Class Type</th><th>Tweets</th><th>Tweets</th></tr>
<tr><td rowspan="3">Modern MGT (MMGT)</td><td>GPT-2</td><td>3839</td><td rowspan="3">5839</td></tr>
<tr><td>GPT-3</td><td>1676</td></tr>
<tr><td>ChatGPT</td><td>493</td></tr>
<tr><td>Traditional MGT (TMGT)</td><td>RNN
Others</td><td>2746
2280</td><td>5026</td></tr>
</table>

EXPLORATORY DATA ANALYSIS

- The **analysis focused** on identifying **distinguishing characteristics between**
  - Machine-Generated Text (**MGT**) and Human-Written Text (**HWT**)
  - Modern Machine-generated Text (**MMGT**) and Traditional Machine-Generated Text (**TMGT**).
- Findings from our analysis are directly relevant to **addressing RQ1** and **RQ2**.

RQ1) What are the distinguishing characteristics between **MGT** and **HWT** on Twitter?

RQ2) What are the distinguishing characteristics between **MMGT** and **TMGT** on Twitter?

EXPLORATORY DATA ANALYSIS

Linguistic Analysis

- **Tweet Indicators Analysis**
  - focused on Twitter-specific components like user mentions, URLs, and hashtags.
- **Part-of-speech (POS) Analysis**
  - involves categorizing words into grammatical classes, such as nouns, verbs, and adjectives.
- **Named-entity-recognition (NER) Analysis**
  - involves categorizing words into named entities, such as people, organizations, locations and dates.

EXPLORATORY DATA ANALYSIS

Linguistic Analysis

- **Text perplexity Analysis**
    - To measure the level of unpredictability or uncertainty in language models' predictions.
    - Lower perplexity indicates that the language model is more confident and accurate in predicting the next word, whereas higher perplexity implies poorer predictive performance.
    - We employed the GPT-2 model for this task.

EXPLORATORY DATA ANALYSIS

Linguistic Analysis

- **Linguistic Feature Importance Analysis**
  - to identify relevant linguistic features for our bot detection classifier

Sentiment Analysis

- To determine the underlying sentiment or emotion expressed in the text
  - whether it is positive, negative, or neutral.

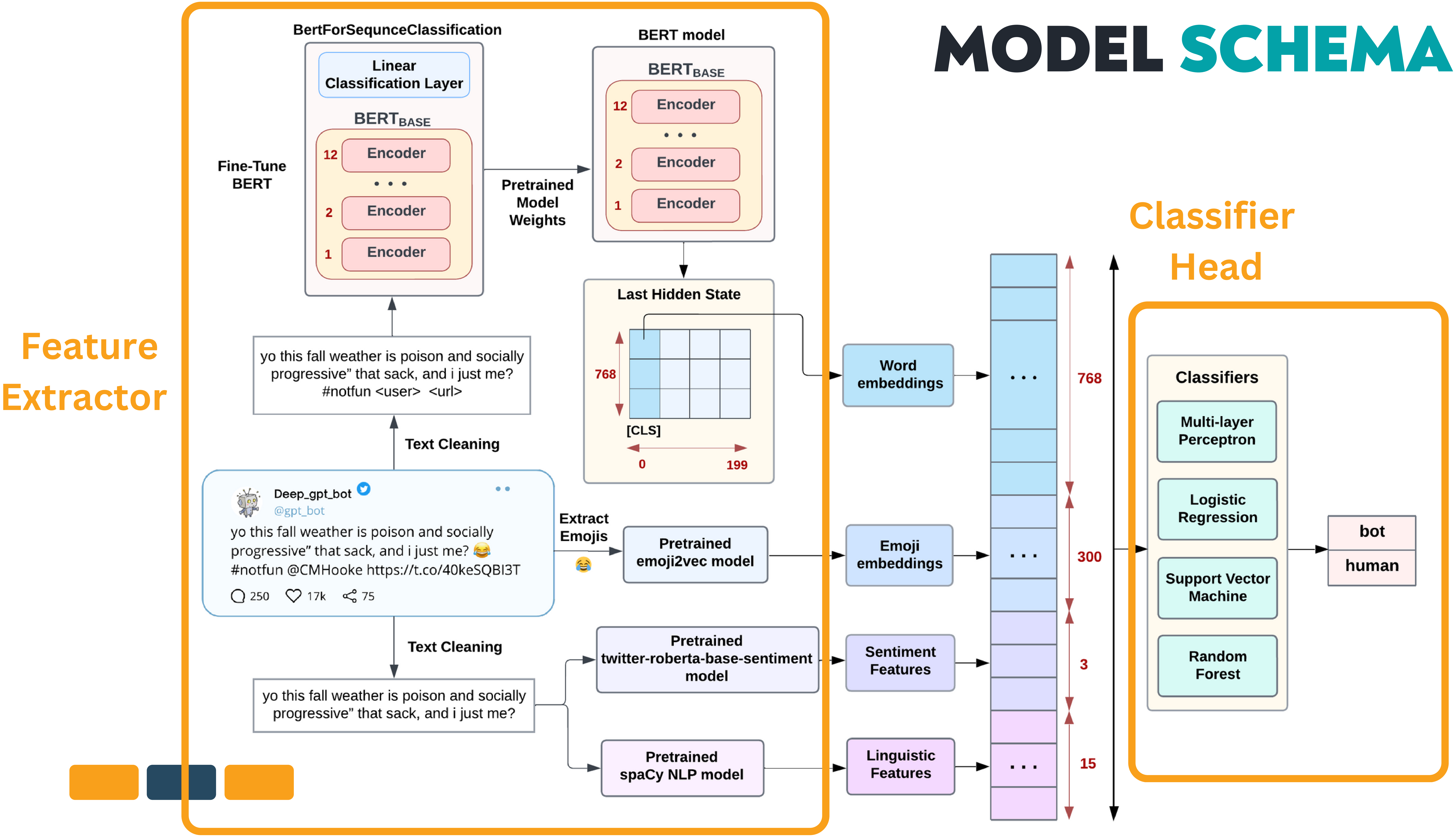
MODELING EXPERIMENTS

- **To develop a discriminator that classifies** machine-generated text **(MGT)** and human-written text **(HWT)**

- **To address RQ3.**

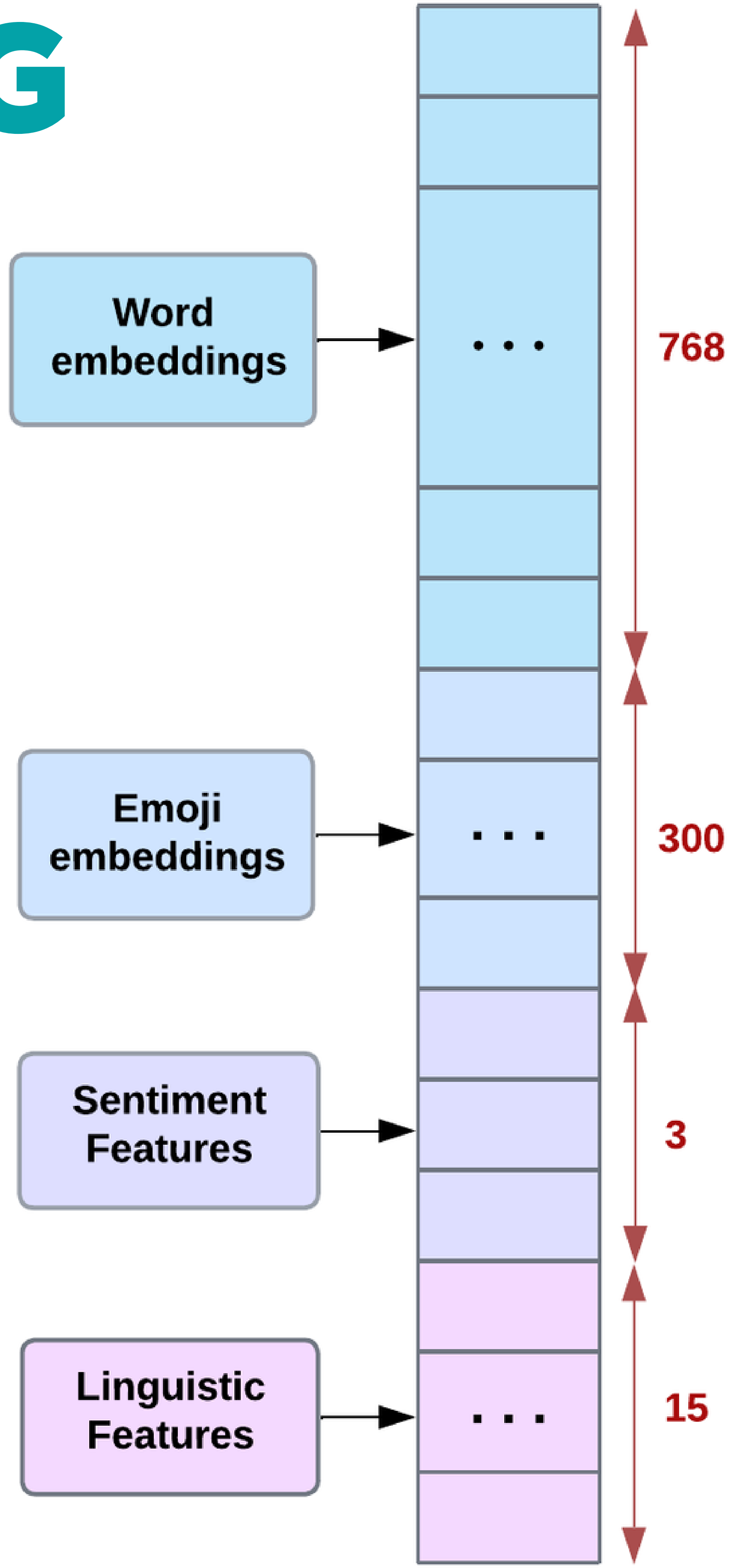
RQ3) To what extent does incorporating linguistic features, sentiment features, and emojis embeddings alongside semantic word embeddings enhance the model's ability to accurately classify MGT and HWT on Twitter?

MODEL SCHEMA



FEATURE ENGINEERING

- **A. Semantic features**
  - Fine-tuned BERT embeddings
    - 768 - dimensions
- **B. Emoji features**
  - emoji2vec pretrained embeddings
    - 300 - dimensions
- **C. Sentiment features**
  - 3 features
- **D. Linguistic features**
  - 15 features



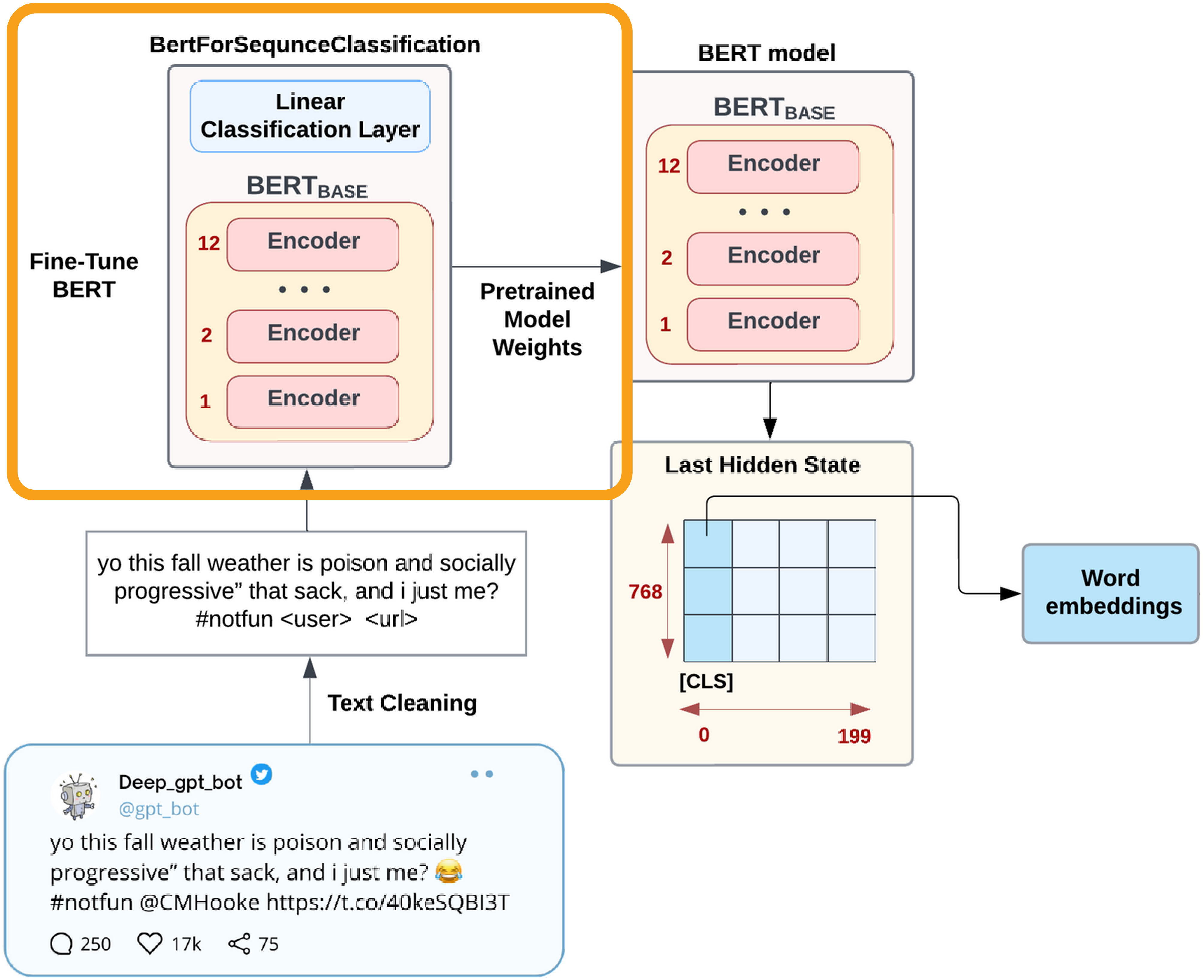
METHODOLOGY

FEATURE ENGINEERING

A. Semantic Features - Fine-tuned BERT embeddings

First step - Fine-tuning BERT

- **fine-tuned a pre-trained BERT model** (Bidirectional Encoder Representations from Transformers)
- utilized BertForSequenceClassification interface
- BERT model variant: BERT-BASE(cased) model
- The **weights of the fine-tuned model were saved**.



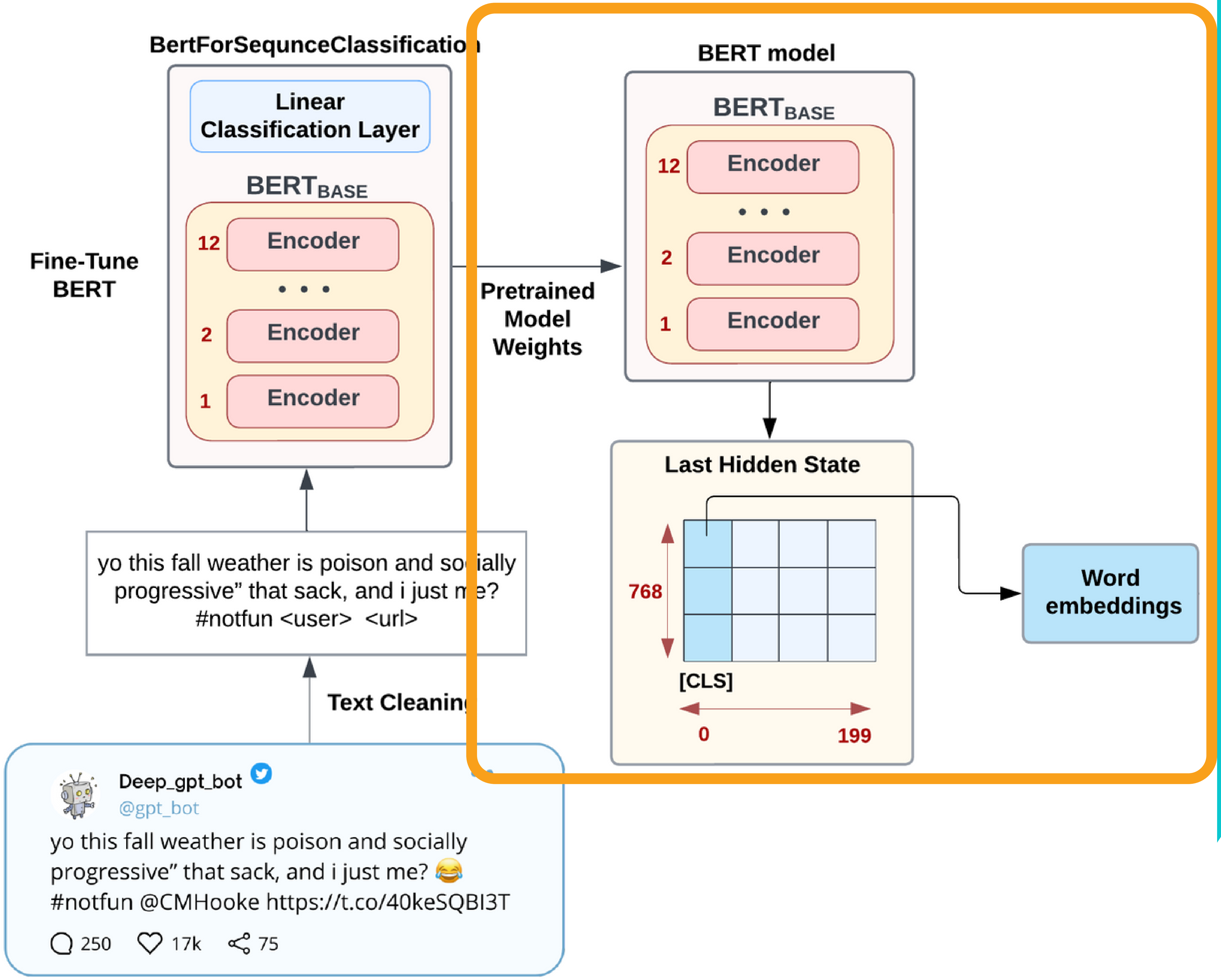
METHODOLOGY

FEATURE ENGINEERING

A. Semantic Features - Fine-tuned BERT embeddings

Second step - Extract BERT Embeddings

- utilized the saved fine-tuned model's weights on the bare BERT model
- extracted **768-dimensional embeddings** from the **last hidden state** of the model corresponding to the **[CLS] token**.
- The [CLS] token acts as a summary representation that encapsulates the semantic understanding of the entire tweet.

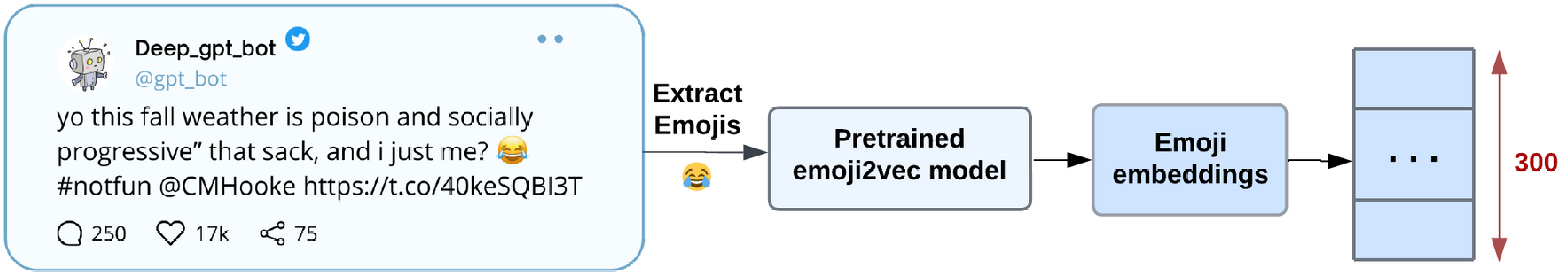


METHODOLOGY

FEATURE ENGINEERING

B. Emoji Features - emoji2vec embeddings

- Original BERT model lacks representation for emoji tokens in its vocabulary
- To address this gap and incorporate emoji features into our model, we turned to emoji2vec.
- **emoji2vec** (Eisner et al., 2016)
  - is a pre-trained embedding model that assigns 300-dimensional vectors to all Unicode emojis



FEATURE ENGINEERING

C. Sentiment Features

- utilized a **pre-trained RoBERTa sentimenet classificaion model**
  - specifically the "twitter-roberta-base-sentiment-latest" model from Hugging Face library

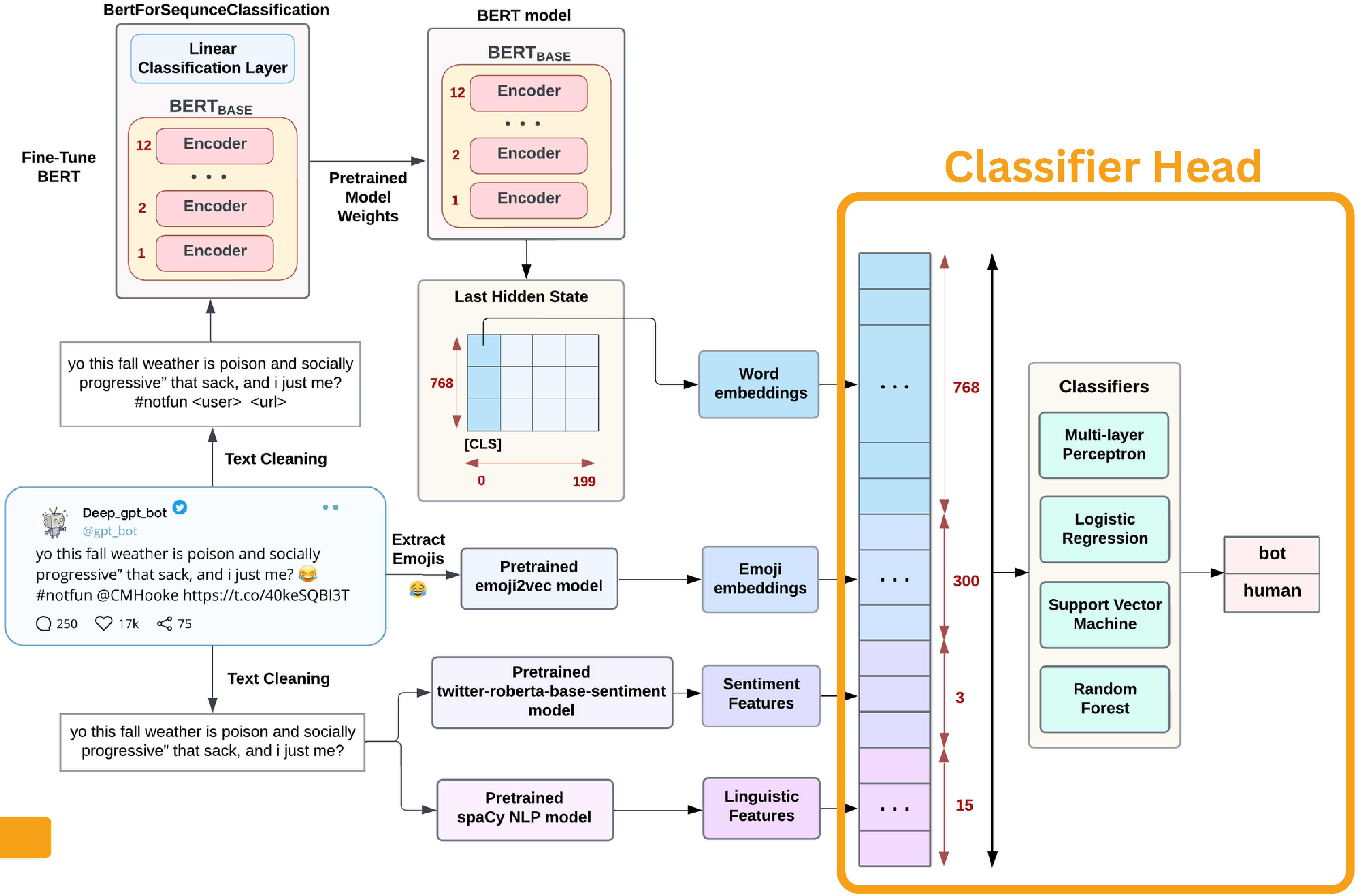
Sentiment Features		
POS	Positive sentiment strength.	Floating Point; [0, 1]
NEG	Negative sentiment strength.	Floating Point; [0, 1]
NEU	Neutral sentiment strength.	Floating Point; [0, 1]

FEATURE ENGINEERING

C. Linguistic Features

Features	Description	Data Type & Domain
Tweet Indicators		
URL	Indicator of whether there are **URLs** present within the tweet.	Boolean; {0, 1}
Mentions	Indicator of whether there are **user mentions** present within the tweet.	Boolean; {0, 1}
Part of Speech (POS)		
NOUN	Ratio of **nouns** to the total number of tokens in a tweet.	Floating Point; [0, 1]
VERB	Ratio of **verb** to the total number of tokens in a tweet.	Floating Point; [0, 1]
PRON	Ratio of **pronoun** to the total number of tokens in a tweet.	Floating Point; [0, 1]
PUNCT	Ratio of **punctuation** to the total number of tokens in a tweet.	Floating Point; [0, 1]
PROPN	Ratio of **proper noun** to the total number of tokens in a tweet.	Floating Point; [0, 1]
ADP	Ratio of **adposition** to the total number of tokens in a tweet.	Floating Point; [0, 1]
DET	Ratio of **determiner** to the total number of tokens in a tweet.	Floating Point; [0, 1]
ADJ	Ratio of **adjective** to the total number of tokens in a tweet.	Floating Point; [0, 1]
ADV	Ratio of **adverb** to the total number of tokens in a tweet.	Floating Point; [0, 1]
AUX	Ratio of **auxiliary** to the total number of tokens in a tweet.	Floating Point; [0, 1]
PART	Ratio of **particle** to the total number of tokens in a tweet.	Floating Point; [0, 1]
CCONJ	Ratio of **coordinating conjunction** to the total number of tokens in a tweet.	Floating Point; [0, 1]
Perplexity		
PPL	**Perplexity Score** based on GPT2 model (log-transformed & scaled).	Floating Point; [0, 1]

MODEL SCHEMA

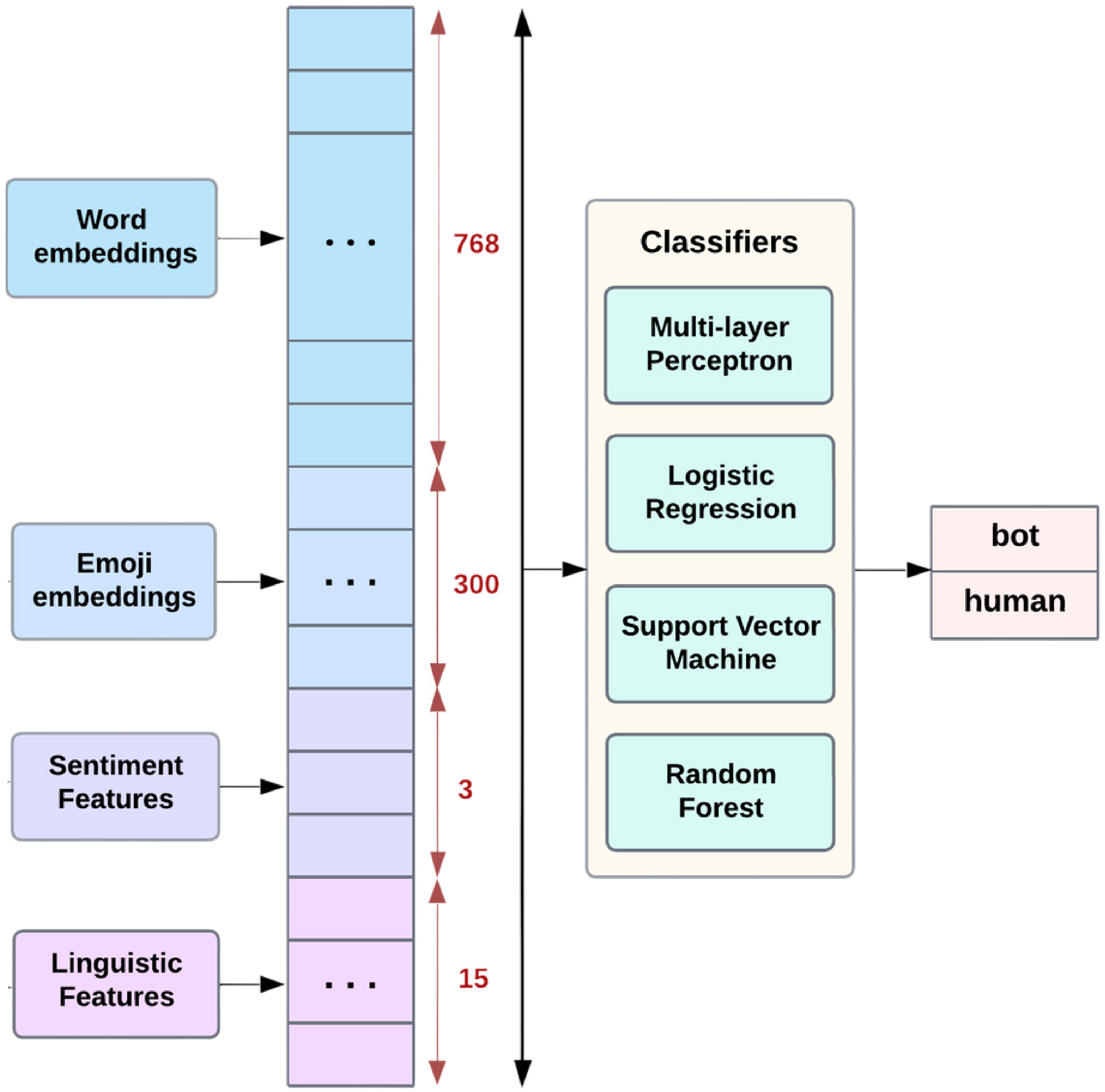


MODEL DESCRIPTION

The Classifier Head

- Deep learning model
  - Multi-Layer Perceptron (MLP)

- Shallow machine learning models
  - Logistic Regression (LR)
  - Support Vector Machines (SVC)
  - Random Forest (RF)



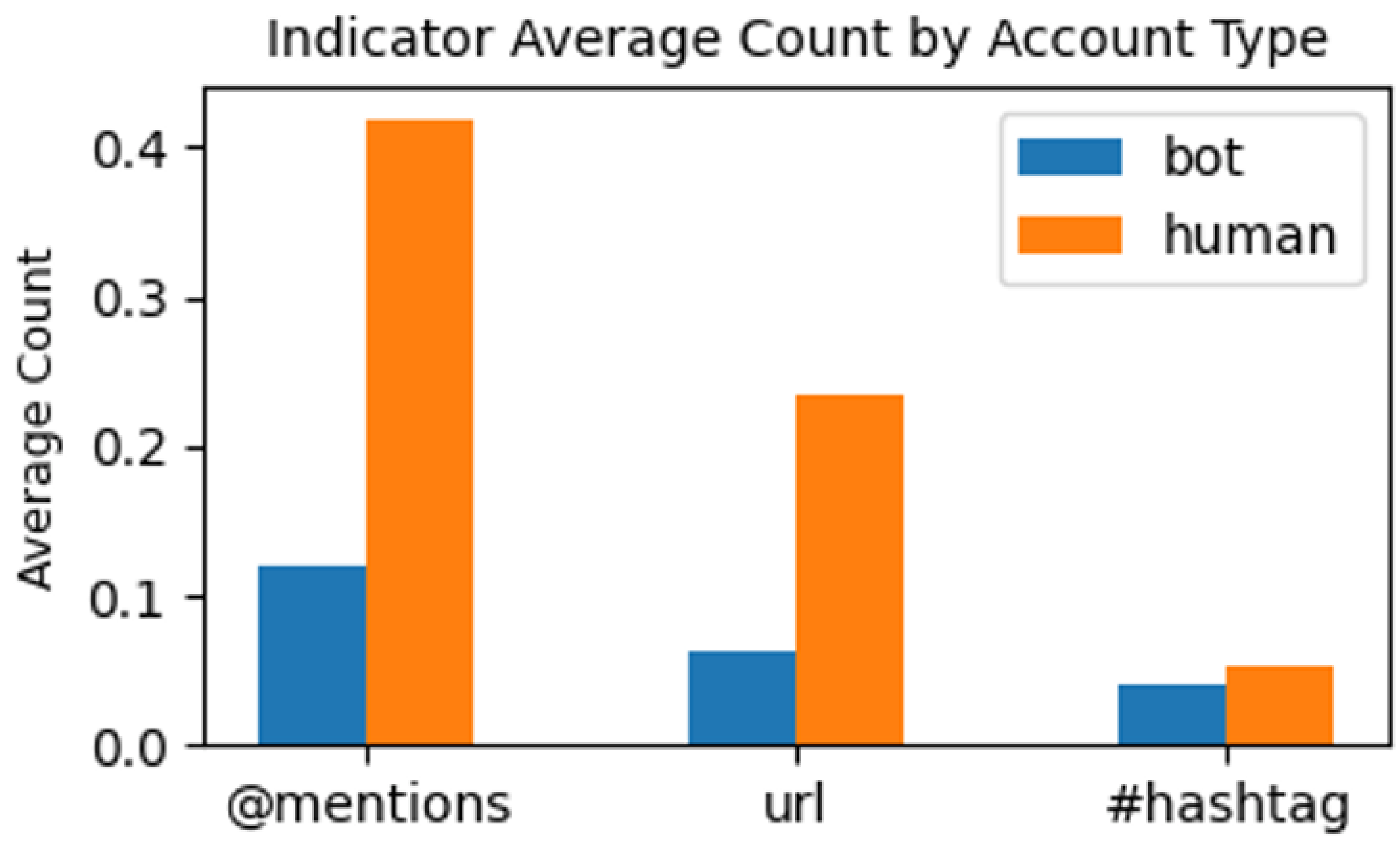
RESULTS & DISCUSSION

RESULTS OF EXPLORATORY DATA ANALYSIS

RQ1) What are the distinguishing characteristics between **MGT** and **HWT** on Twitter?

Tweets Indicators Feature

- **HWT displayed a significantly higher frequency of user mentions and URLs compared to MGT**.
  - This suggests that human users are more inclined to engage with others by mentioning them and sharing external links in their tweets.



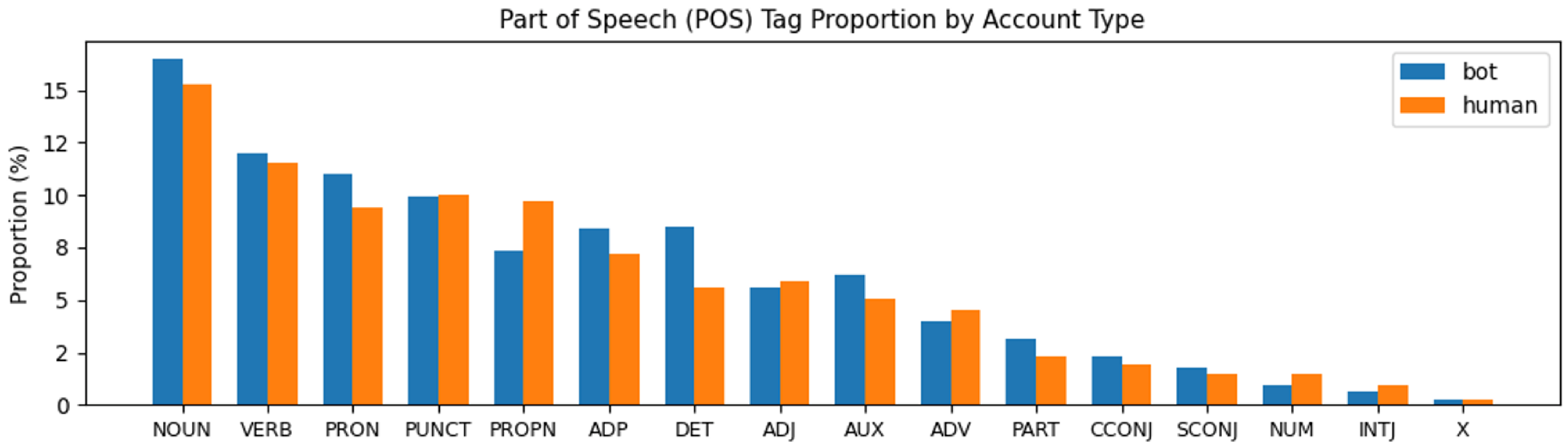
RESULTS & DISCUSSION

RESULTS OF EXPLORATORY DATA ANALYSIS

RQ1) What are the distinguishing characteristics between **MGT** and **HWT** on Twitter?

Part-of-speech (POS) Analysis

- **MGT** utilize a **higher frequency of words** related to **noun, pronoun, determiner**, and **adposition**, but **fewer words** related to **proper noun** and **adverb** as compared to HWT.



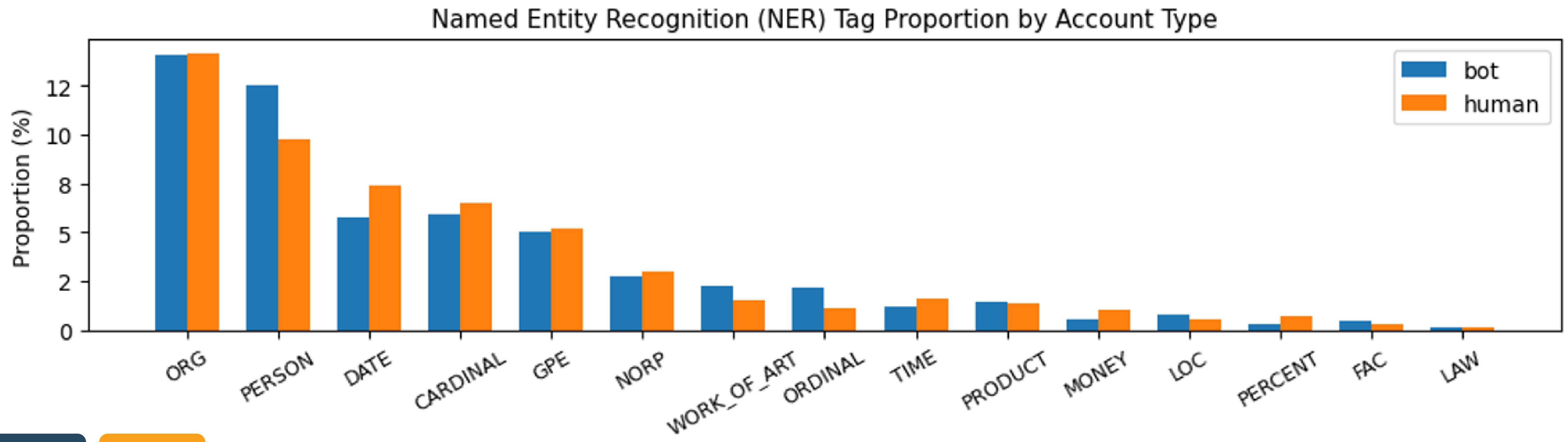
RESULTS & DISCUSSION

RESULTS OF EXPLORATORY DATA ANALYSIS

RQ1) What are the distinguishing characteristics between **MGT** and **HWT** on Twitter?

Named-entity-recognition (NER) Analysis

- **MGT demonstrates a higher frequency of words related to PERSON entities**
  - possibly due to the models' exposure to social media content where individuals' names and mentions are prevalent
- **MGT exhibits a lower frequency of words related to DATE entities**
  - indicating a preference for generating content that is less time-dependent.



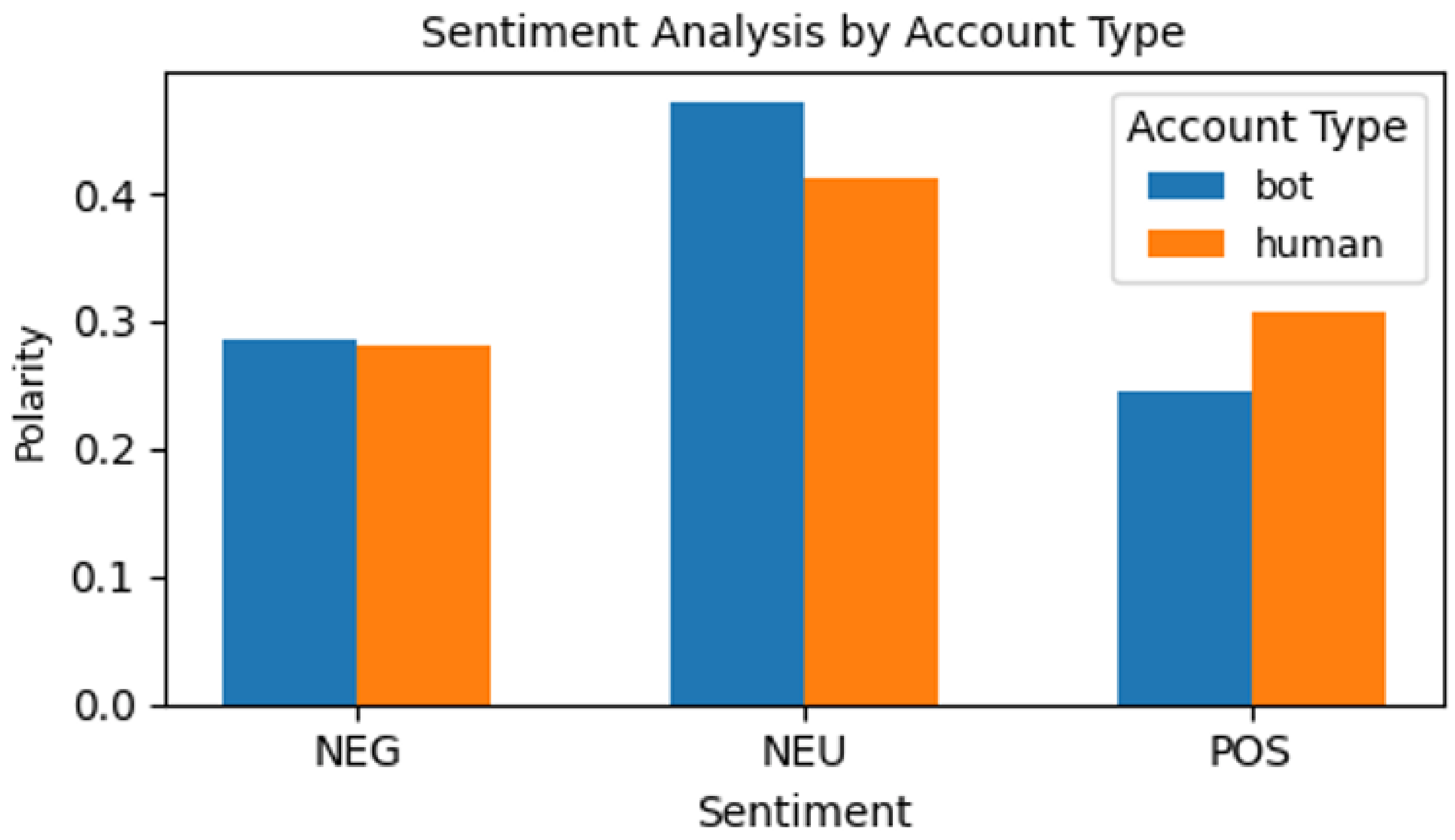
RESULTS & DISCUSSION

RESULTS OF EXPLORATORY DATA ANALYSIS

RQ1) What are the distinguishing characteristics between **MGT** and **HWT** on Twitter?

Sentiment Analysis

- **MGT exhibits higher proportion of neutral sentiment but a lower proportion of positive sentiment compared to HWT**
  - indicates that text generative models produce content with a neutral tone, avoiding strong opinions or emotions.

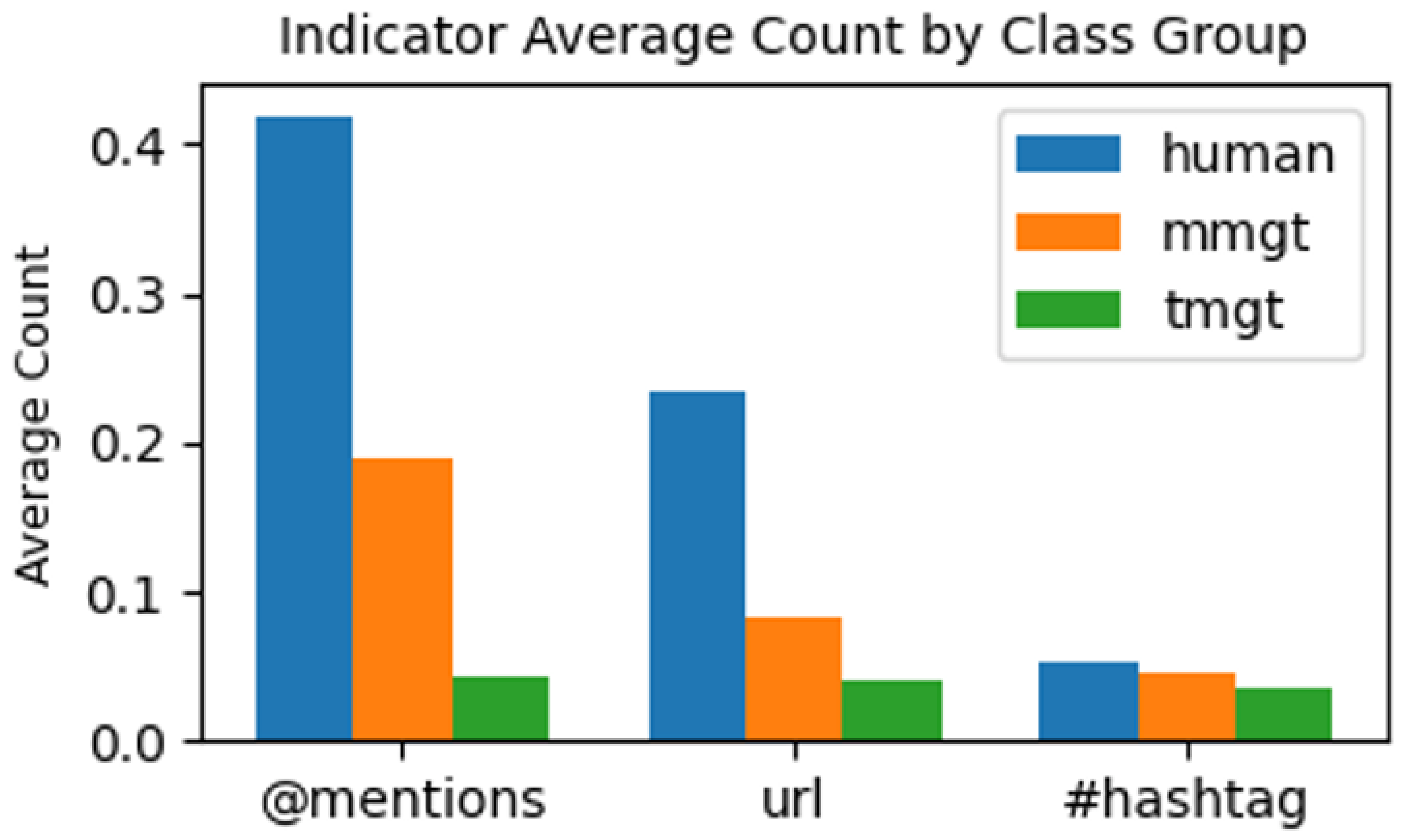


RESULTS OF EXPLORATORY DATA ANALYSIS

RQ2) What are the distinguishing characteristics between **MMGT** and **TMGT** on Twitter?

Tweets Indicators Feature

- **MMGT demonstrates significantly higher numbers of user mentions and URLs compared to TMGT**
  - This suggests that modern generative models have made advancements in mimicking human-like behavior by engaging with other users and sharing external content.

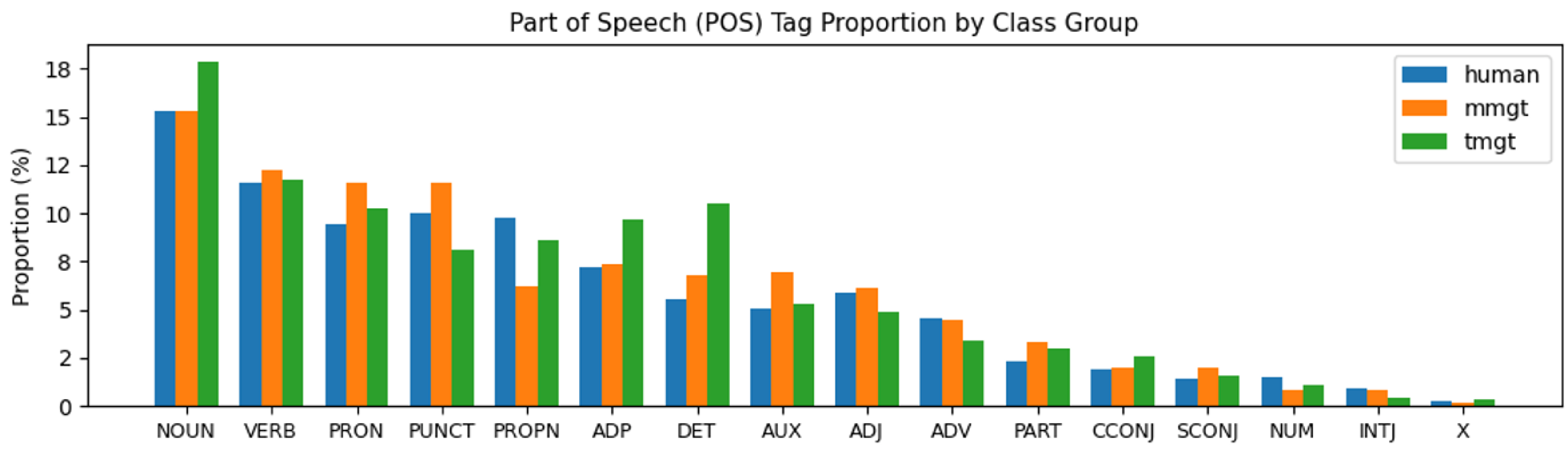


RESULTS OF EXPLORATORY DATA ANALYSIS

RQ2) What are the distinguishing characteristics between **MMGT** and **TMGT** on Twitter?

Part-of-speech (POS) Analysis

- **MMGT** displays **lower frequencies** of **noun, pronoun, determiner, adposition, proper noun** while using **more punctuation, pronoun** and **auxiliary**.
  - This suggests that modern generative models have advanced in capturing and utilizing a wider range of grammatical structures and linguistic patterns.



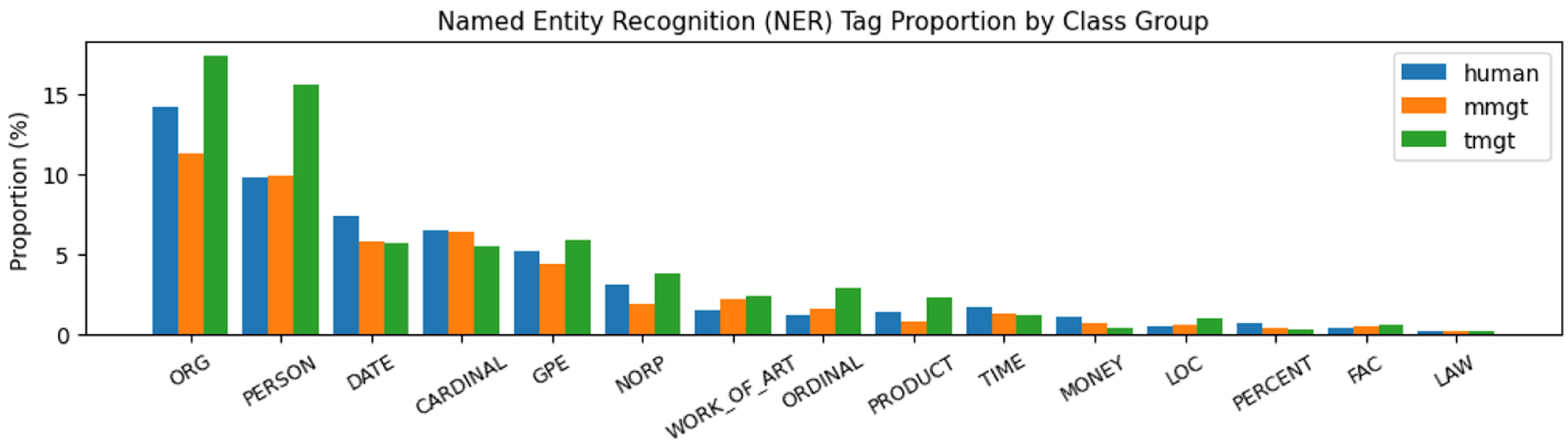
RESULTS & DISCUSSION

RESULTS OF EXPLORATORY DATA ANALYSIS

RQ2) What are the distinguishing characteristics between **MMGT** and **TMGT** on Twitter?

Named-entity-recognition (NER) Analysis

- **MMGT exhibits a lower frequency of words related to organizations and person entities compared to TMGT**
  - This indicates that modern generative models generate content with a reduced emphasis on specific organizations and individuals, aligning more closely with general patterns and topics.



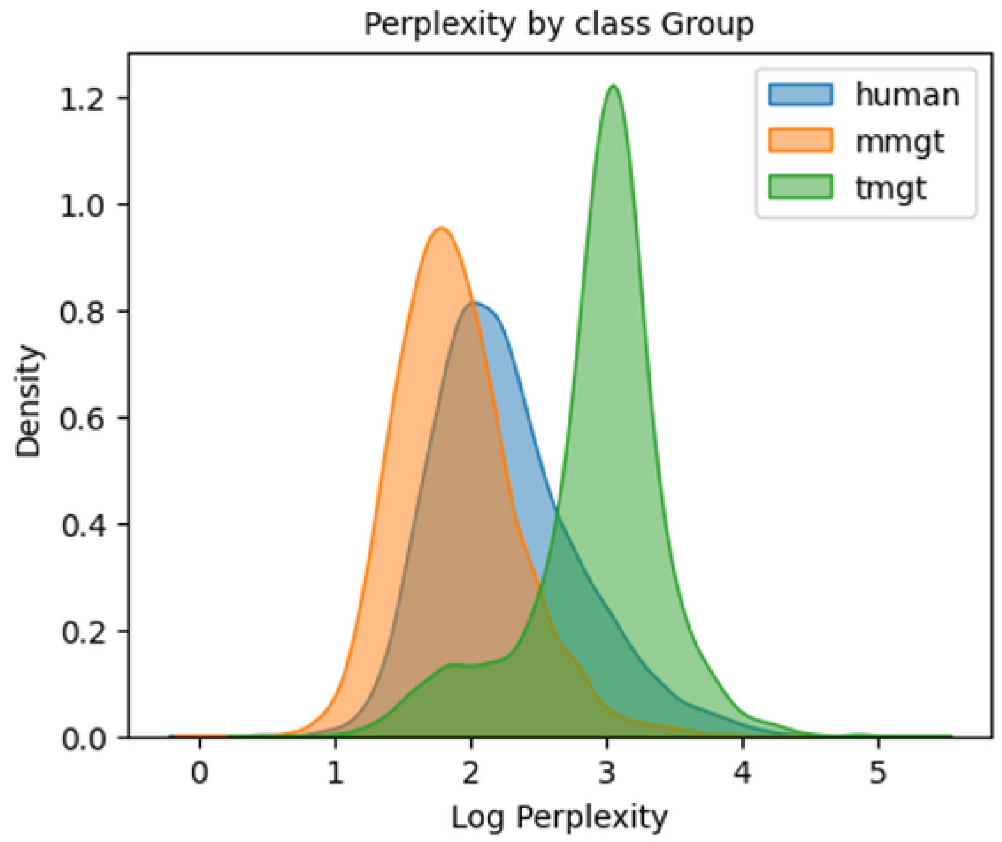
RESULTS & DISCUSSION

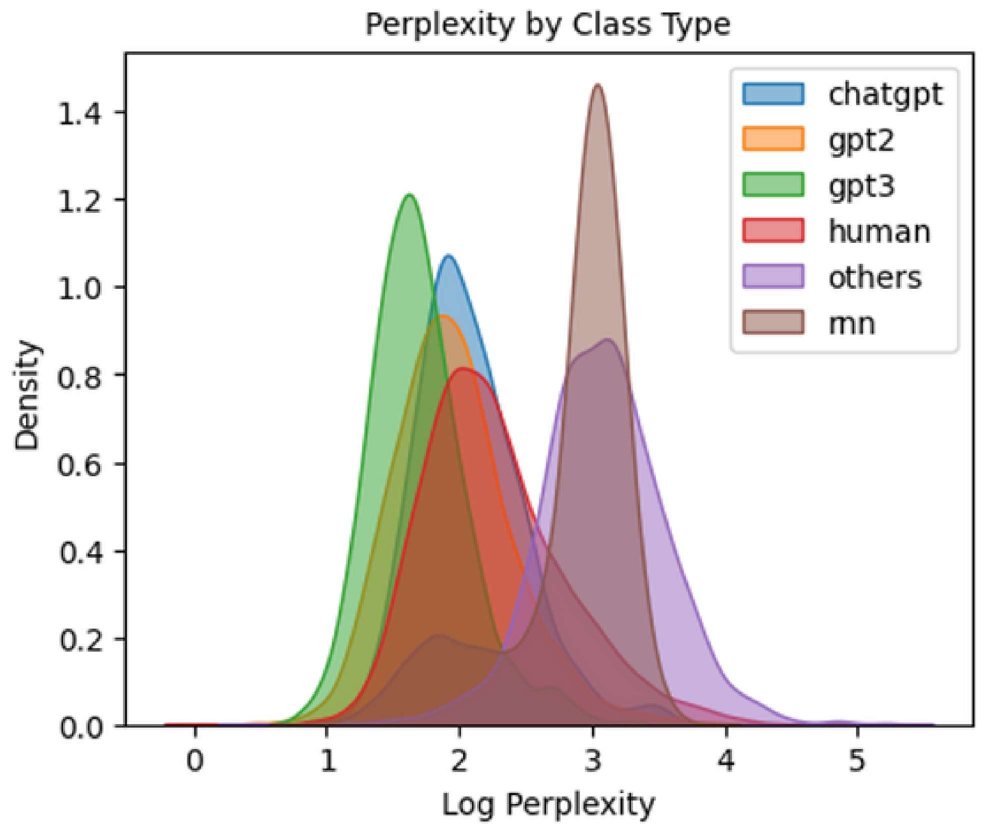
RESULTS OF EXPLORATORY DATA ANALYSIS

RQ2) What are the distinguishing characteristics between **MMGT** and **TMGT** on Twitter?

Text Perplexity Analysis

- **MMGT exhibits relatively low perplexity compared to TMGT**
  - Modern generative text models, such as GPT-2, GPT-3, and ChatGPT, capture common patterns from their training data, allowing them to replicate such patterns effectively.
  - Consequently, when computing text perplexity using the GPT-2 model, it becomes less perplexed by text generated by similar modern generative text models.





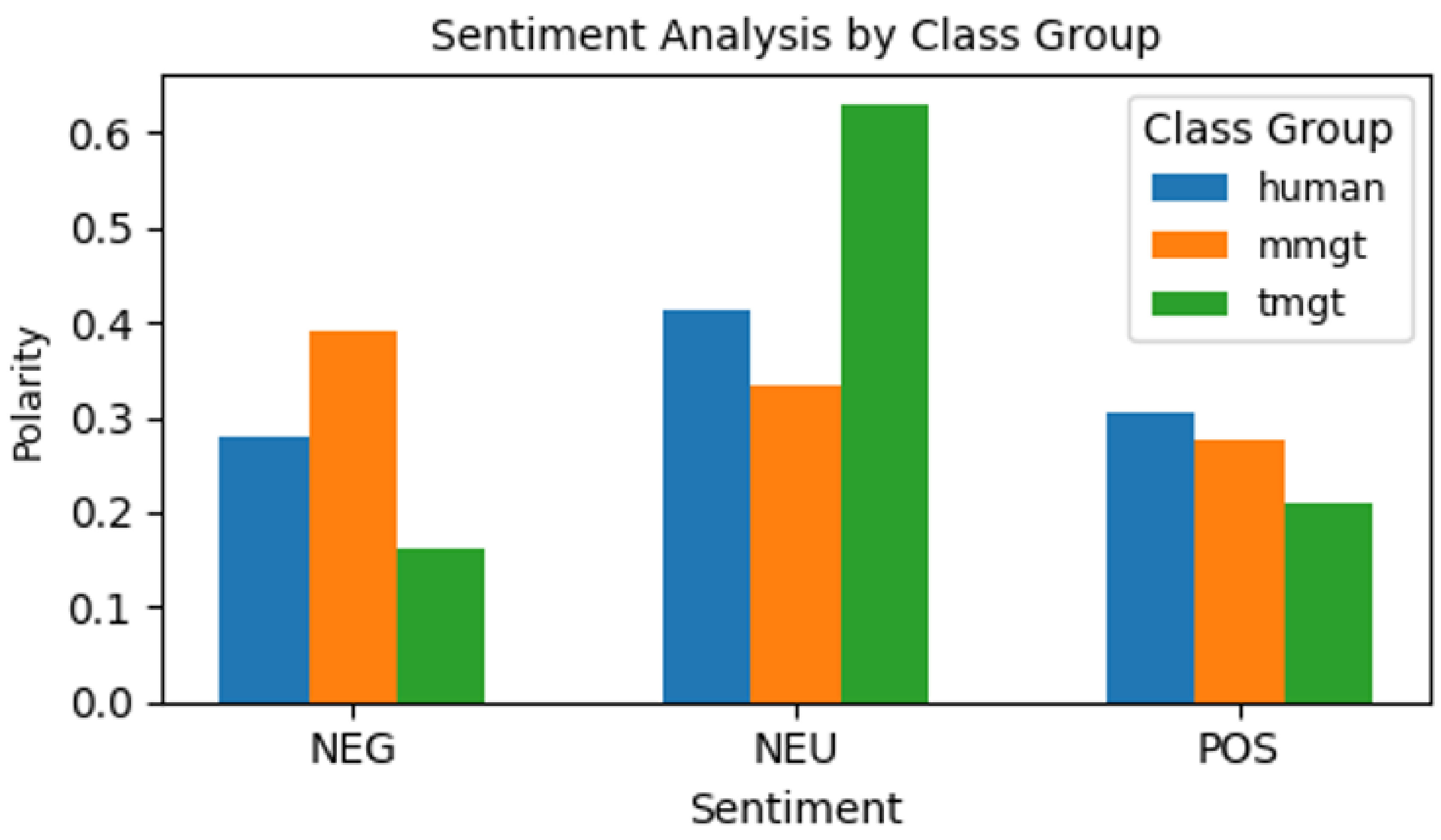
RESULTS & DISCUSSION

RESULTS OF EXPLORATORY DATA ANALYSIS

RQ2) What are the distinguishing characteristics between **MMGT** and **TMGT** on Twitter?

Sentiment Analysis

- **TMGT shows a significant increase in neutral tweets and a reduction in negative and positive tweets relative to MMGT**
  - indicates that traditional generative models may have limitations in generating content with nuanced sentiment expressions



RESULTS OF MODELING EXPERIMENTS

RQ3) To what extent does incorporating linguistic features, sentiment features, and emojis embeddings alongside semantic word embeddings enhance the model's ability to accurately classify MGT and HWT on Twitter?

- The incorporation of **BERT embeddings with additional features**, specifically **emojis and linguistic features, consistently outperforms using BERT embeddings alone**
  - increased accuracy ranging from 0% to 0.6%.
- **Sentiment features did not contribute significantly** to the performance improvement with most classifiers.
- **Best performing model**:
  - Fine-tuned BERT embeddings in combination with emoji and linguistic features, while employing the MLP classifier.
  - Accuracy rate: 88.3%.

RESULTS OF MODELING EXPERIMENTS

Classifier	Features	Human			Bot			Globally
		Precision	Recall	F1	Precision	Recall	F1	Accuracy
LR	BERT	0.875	0.864	0.869	0.866	0.877	0.871	0.870
	BERT + Emoji	0.877	0.864	0.870	0.866	0.879	0.872	0.871
	BERT + Sent	0.875	0.864	0.869	0.866	0.877	0.871	0.870
	BERT + Ling	0.875	0.867	0.871	0.869	0.877	0.873	**0.872**
	BERT + Emoji + Sent	0.876	0.862	0.869	0.864	0.879	0.871	0.870
	BERT + Emoji + Ling	0.875	0.867	0.871	0.869	0.877	0.873	**0.872**
	BERT + Sent + Ling	0.875	0.866	0.870	0.867	0.877	0.872	0.871
	BERT + Emoji + Sent + Ling	0.875	0.866	0.870	0.867	0.877	0.872	0.871
SVC	BERT	0.869	0.858	0.864	0.860	0.871	0.866	0.865
	BERT + Emoji	0.867	0.862	0.864	0.863	0.868	0.865	0.865
	BERT + Sent	0.870	0.862	0.866	0.863	0.871	0.867	0.867
	BERT + Ling	0.871	0.867	0.869	0.868	0.871	0.870	**0.869**
	BERT + Emoji + Sent	0.869	0.864	0.866	0.865	0.869	0.867	0.867
	BERT + Emoji + Ling	0.869	0.864	0.866	0.865	0.869	0.867	0.867
	BERT + Sent + Ling	0.871	0.867	0.869	0.868	0.871	0.870	**0.869**
	BERT + Emoji + Sent + Ling	0.865	0.864	0.865	0.864	0.866	0.865	0.865

Classifier	Features	Human			Bot			Globally
		Precision	Recall	F1	Precision	Recall	F1	Accuracy
RF	BERT	0.865	0.888	0.876	0.885	0.862	0.873	0.875
	BERT + Emoji	0.870	0.888	0.879	0.886	0.868	0.877	0.878
	BERT + Sent	0.870	0.884	0.877	0.882	0.868	0.875	0.876
	BERT + Ling	0.873	0.890	0.881	0.888	0.871	0.879	0.880
	BERT + Emoji + Sent	0.868	0.888	0.878	0.885	0.866	0.875	0.877
	BERT + Emoji + Ling	0.870	0.890	0.880	0.887	0.868	0.877	0.879
	BERT + Sent + Ling	0.873	0.890	0.881	0.888	0.871	0.879	0.880
	BERT + Emoji + Sent + Ling	0.875	0.890	0.882	0.888	0.873	0.880	**0.881**
MLP	BERT	0.872	0.89	0.881	0.887	0.869	0.878	0.879
	BERT + Emoji	0.878	0.888	0.883	0.887	0.877	0.882	0.882
	BERT + Sent	0.880	0.876	0.878	0.877	0.880	0.879	0.879
	BERT + Ling	0.874	0.893	0.883	0.891	0.871	0.881	0.882
	BERT + Emoji + Sent	0.877	0.885	0.880	0.884	0.876	0.879	0.880
	BERT + Emoji + Ling	0.886	0.879	0.882	0.880	0.887	0.883	**0.883**
	BERT + Sent + Ling	0.875	0.886	0.881	0.885	0.874	0.879	0.880
	BERT + Emoji + Sent + Ling	0.882	0.880	0.881	0.881	0.882	0.882	0.881

Note: BERT refers to BERT embeddings obtained after the fine-tuning phase, while Emoji represents emoji2vec features. Sent represents 3 sentiment features, and Ling represents 15 linguistic features. Bold scores indicate the highest accuracy score for each classifier. The highlighted row represents the combination that yielded the highest F1-score and accuracy on the test set among all feature and classifier combinations.

COMPARISON TO RESULTS IN THE LITERATURE

Related Research	Approach Summary	Features	Performance	Performance
		Dataset:	TweepFake (Fagni et al., 2021)	Enhanced TweepFake
Fagni et al. (2021)	Fine-tuned RoBERTa model	Semantic Embeddings	Acc: 89.6% $F1_{Bot}$: 89.7% $F1_{Human}$: 89.5%	Acc: 87.6% $F1_{Bot}$: 87.6% $F1_{Human}$: 87.6%
Fagni et al. (2021)	Fine-tuned BERT model	Semantic Embeddings	Acc: 89.1% $F1_{Bot}$: 89.2% $F1_{Human}$: 89.0%	Acc: 86.8% $F1_{Bot}$: 87.0% $F1_{Human}$: 86.7%
Saravani et al. (2021)	BERT (word embeddings) + BILSTM (capture temporal relations) + NeXtVLAD (parametric pooling area)	Semantic Embeddings	Acc: 92% $F1_{Bot}$: 92% $F1_{Human}$: 92%	-
Gambini et al. (2020)	Fine-tuned GPT-2-based classifier	Semantic Embeddings	Acc: 91%	-
Tesfagergish et al. (2021)	Fine-tuned RoBERTa (word embeddings) + Hierarchical Attention Network (classifier)	Semantic Embeddings	Acc: 89.7% F1: 85.5%	-
Our Proposed Model	Fine-tuned BERT(word embeddings) + emoji2vec embeddings + linguistic features + Multi-Layer Perceptron (classifier)	Semantic Embeddings, Emoji embeddings, Linguistic features	-	Acc: 88.3% $F1_{Bot}$: 88.3% $F1_{Human}$: 88.2%

CONCLUSION

CONCLUSION

- Our research contributes to the field of deepfake text detection by demonstrating that incorporating semantic text features with supplementary features like emoji and linguistic features enhances the model's ability to detect deepfake text.
- We provided insights in distinct characteristics of MGT, including differences in engagement behavior, linguistic patterns, named entities, sentiment expressions, and text perplexity.
- We also enhanced the TweepFake dataset by including deepfake tweets from the latest text generative models.

CONCLUSION

LIMITATIONS

- **Our research dataset differs from the benchmark dataset** (Original TweepFake dataset) used in the study by Fagni et al. (2021)
  - **making direct comparisons challenging**

- Since our **research focuses on short texts**, our findings may have **limited generalizability to longer textual content**.
  - It's essential to consider potential variations in performance for longer texts.

FUTURE WORKS

- To **investigate** the **generalization capability** of our proposed detector by **evaluating its performance** in **detecting deepfake text** from **previously unseen accounts**
    - To assess the model's robustness and its applicability beyond the training dataset in real-world scenarios.

- To **explore other variations of Transformer models**
    - e.g. RoBERTa, DistilBERT, XLNET

- To **incorporate more sophisticated features** and **conducting in-depth feature importance analysis**,
    - such as utilizing SHAP (SHapley Additive exPlanations), to gain deeper insights into the discriminative power of individual features.

THANK YOU

For Your Attention

SUNWAY UNIVERSITY