Human vs Machine Generated Text

Team03 Members

Yu-Hsuan Lin
Shibo Chen
Yueyang Xu
Jessica James

TABLE OF CONTENTS

Introduction

- Development of a robust model to distinguish between human written and machine generated text.
- Chosen Dataset - Twitter data in diverse topics such as Fifa, election and Game of Thrones
- Exploration of distinct characteristics and patterns in language and content.

Motivation

- Combating misinformation and enhancing trustworthiness in social media platforms, particularly twitter.
- Leveraging Machine Learning techniques to automate the detection to avoid consequences in popular areas - elections, sports events, TV shows.

Data Description

- Datasets chosen for the study - US Election 2020, FIFA World Cup 2022, Game of Thrones S8 from Kaggle
- Randomly extracted 10,000 rows from each of the three datasets.
- Total of 30,000 rows and 10 columns.
- Dataframe - Cleaned Tweets, Topics, emoji count, profile count, URL count, token length, cleaned tweet length, tokenized clean tweet, Human Generated(Label)

Data Preprocessing

- Removal of Punctuation marks
- Removal of @profiles, emojis, and URLs
- Count the number of @profiles, emojis, and URLs
- Convert to lowercase
- Lemmatization using SpaCy
- Word Tokenization
- Select tweets in English using Language Detection model in FastText to capture informal texts
- Tweet length
- Keep stopwords in the text data for higher accuracy after testing
- randomly sample 10,000 rows to ensure representative sampling

Data Preprocessing Example

"What an episode🙌 We can't wait to see what else Game of Thrones® has in store While we're waiting here's some suggestions on how to play #GOT in Northern Ireland 👉https//tco/RWF4MA2961\n\n 📍 The Dark Hedges County Antrim\n📸 https//tco/lrd2T5vvYy https//tco/IF45j9e73l"

'what an episode we can t wait to see what else game of throne have in store while we re wait here s some suggestion on how to play get in northern ireland the dark hedge county antrim'

	clean_tweet	Topics	Human_generated	count_url	count_emoji	count_profile	clean_tweet_Length	Tweet_Length	Tweet	tokenize_clean_tweet
7376	what an episode we can t wait to see what else...	Game	1	3	5	0	210	264	What an episode🙌 We can't wait to see what els...	['what', 'an', 'episode' 'we', 'can', 't', 'w..

Characteristics Retrieval

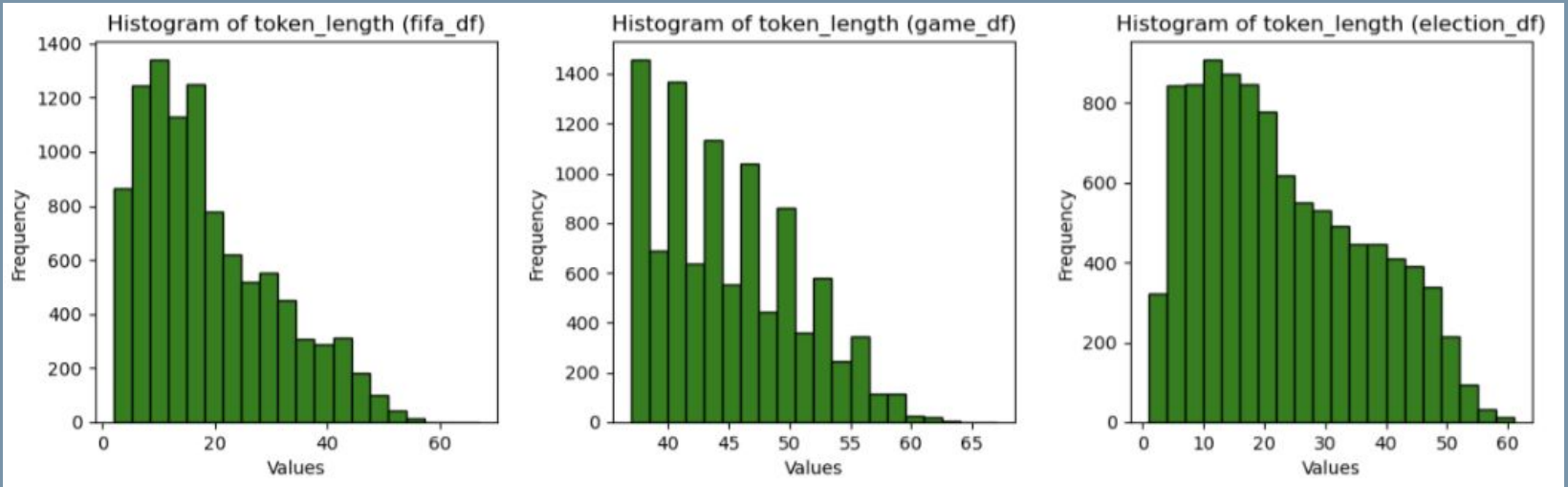


Figure. Distribution of Token Length

Characteristics Retrieval

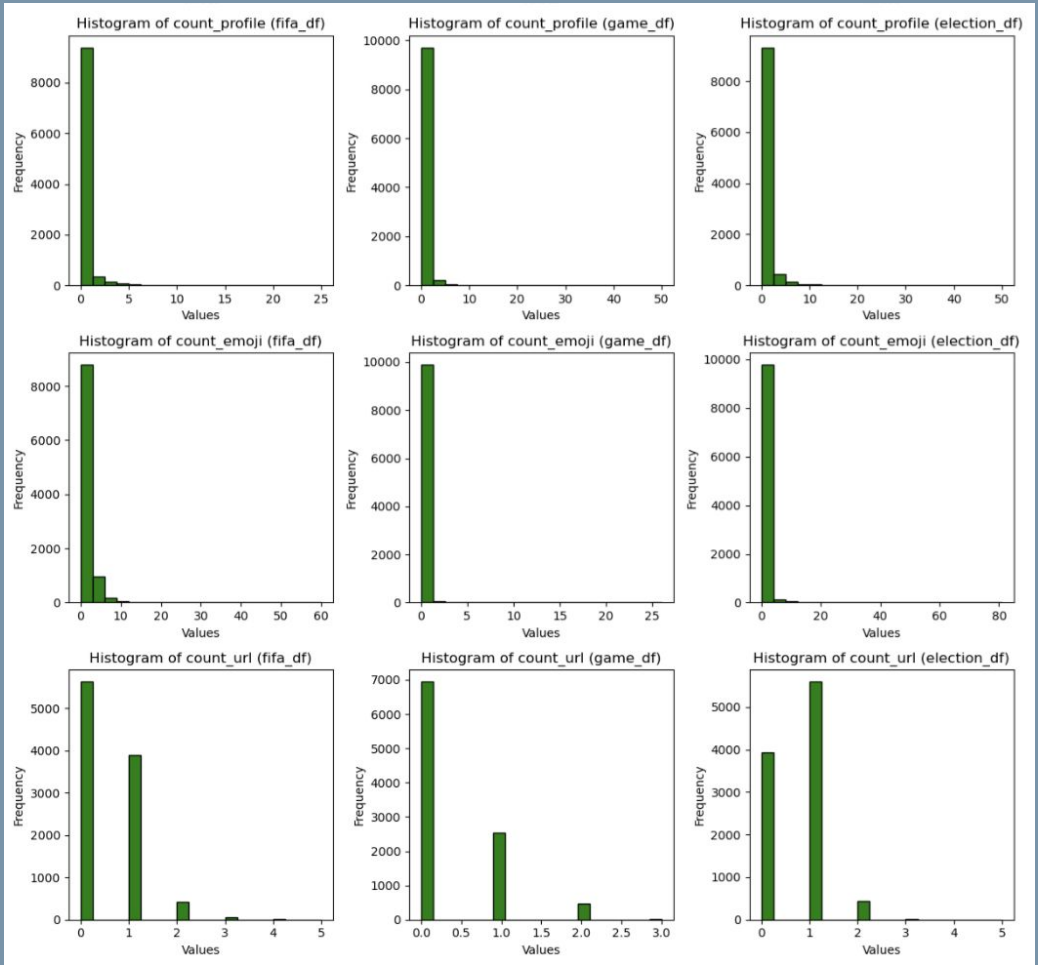


Figure. Distribution of Profile Mentions, Emojis, and URLs in Twitter Data

Characteristics Retrieval

Emojis Count

	Mean	SD	Max
FIFA	**0.86**	2.01	60
Game of Throne	0.053	0.51	26
Election	0.41	1.93	**81**

Profiles Count

	Mean	SD	Max
FIFA	0.3	0.85	**60**
Game of Throne	0.41	1.52	50
Election	**0.76**	2.25	50

URLs Count

	Mean	SD	Max
FIFA	0.49	0.61	5
Game of Throne	0.36	0.58	3
Election	0.56	0.58	5

Characteristics Retrieval

Token Count

	Mean	SD	Max
FIFA	18.71	11.76	67
Game of Throne	**44.8**	5.66	67
Election	22.41	13.65	61

❖ Initial observation

- People exhibit greater enthusiasm and emotional expression in Election and FIFA
- More meaningful language and discussion in Game of Thrones
- No big difference in the distribution of URLs among three datasets, it can be discarded from model training process

❖ Estimation for the Result of Model Training

- When TF-IDF is input, Game of Thrones may have a higher accuracy than others
- When Emoji & profiles is input, Election and FIFA may have higher accuracy than Game of Thrones

Machine Generated TEXT Retrieval

- **Function Purpose**: We defined a function `generate_tweets(prompt)` to send a request to the OpenAI API, requesting the GPT-3.5-turbo model to generate a set of tweets based on a provided prompt. It is designed to automate the process of creating machine-generated content for social media platforms like Twitter.
- **Message Formatting**: The function formats the prompt into a message object, specifying the user's role and the content of the prompt. This formatted message is then sent to the GPT-3.5-turbo model for generating responses.
- **API Interaction**: The function interacts with OpenAI's API using the provided API key. It sends a request to the GPT-3.5-turbo model, specifying the model to use and the prompt provided as user input.
- **Response Handling**: Split generated tweets with the newline character to separate individual tweets; return them as a list of strings, providing the machine-generated tweets for further processing or display. Extract the content of the generated completion using: response.choices[0].message.content.

Prompt Engineering

- **Initial Approach: Encourage creativity and variety**
  - **Sample prompt:**
  - "Write a Twitter post related to the US airline industry from the perspective of an airline passenger. You can focus on key topics like safety measures, customer service experiences, flight delays, sustainability efforts, or recent news in the industry. You can explore different perspectives, sentiments, tones, and natural language patterns. Feel free to ask questions, share personal experiences, or express opinions in your tweets."
- **Observation: Generated tweets differ from human tweets in many aspects**
  - Sentiment diversity
  - Context completeness
  - Tone
  - Spelling

Prompt Engineering (cont)

- **Sample Tweets:**

- **Human:**
  - @VirginAmerica What happened 2 ur vegan food options?! At least say on ur site so i know I won't be able 2 eat anything for next 6 hrs #fail

- **Machine:**
  - Winter weather can cause flight delays but it's essential for airlines to prioritize safety above all else. As passengers we understand the need for caution and appreciate the extra precautions taken during inclement weather #WinterTravelAwarenes

Prompt Engineering (cont)

- **New approach: Paraphrasing**
  - Paraphrase every user tweet in 1:1 ratio
  - Incorporate human tweets as examples within the prompt
  - Emphasize on imitating content and tone
- **Modified prompt:**
  - Mimic the sample tweets for style and content. Contextual meaning (style, content) should be the same as the sample tweets. Use an informal tone. At least 60% of the vocabulary you use must come from the sample tweet:
  - So the writer of Game of Thrones came into my job today I really wanted to press him about Cersei getting crushed by rocks of all ways to die
- **Machine generated tweet:**
  - Can you believe the Game of Thrones writer showed up at my job today? I was so tempted to ask him about Cersei's wild death by rocks like really rocks?

Results

Fifa World Cup 2020 Datasets

Feature\Model	RF(%)	DT(%)	SVM(%)	LR(%)
Clean tweets	72.74	64.84	75.00	76.44
Emoji_profile count	59.00	59.10	51.60	53.67
Clean_tweets_ emojis_profiles	73.80	65.10	60.40	76.90

RF = Random Forest Classifier
DT = Decision Tree
SVM = Support Vector Machine
LR = Logistic Regression

K-fold cross-validation (k=5) used for every model

Table 1. Summary of Categorization Accuracy for Four Models Across Two Features

Results

Game of Thrones S8 Datasets

Feature\Model	RF(%)	DT(%)	SVM(%)	LR(%)
Clean tweets	99.96	99.98	99.95	99.76
Emoji_profile count	54.00	53.99	51.45	53.78
Clean_tweets_ emojis_profiles	99.98	99.98	87.62	99.66

RF = Random Forest Classifier
DT = Decision Tree
SVM = Support Vector Machine
LR = Logistic Regression

K-fold cross-validation (k=5) used for every model

Table 2. Summary of Categorization Accuracy for Four Models Across Two Features

Results

US Election 2020 Datasets

Feature\Model	RF(%)	DT(%)	SVM(%)	LR(%)
Clean tweets	66.41	59.74	71.57	72.08
Emoji_profile count	55.21	55.16	47.83	53.58
Clean_tweets_ emojis_profiles	67.19	60.67	58.25	72.38

RF = Random Forest Classifier
DT = Decision Tree
SVM = Support Vector Machine
LR = Logistic Regression

K-fold cross-validation (k=5) used for every model

Table 3. Summary of Categorization Accuracy for Four Models Across Two Features

Results

Merged Datasets

Feature\Model	RF(%)	DT(%)	SVM(%)	LR(%)
Clean tweets	79.26	74.56	79.71	80.94
Emoji_profile count	49.59	49.62	49.88	49.54
Clean_tweets_ emojis_profiles	49.75	N/A	66.54	49.41

RF = Random Forest Classifier
DT = Decision Tree
SVM = Support Vector Machine
LR = Logistic Regression

K-fold cross-validation (k=5) used for every model

Table 4. Summary of Categorization Accuracy for Four Models Across Two Features

Results

Class	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
Machine_generated	63.00	63.00	71.20	66.80
Human_generated	67.00	67.00	58.60	62.40

Table 5. Summary of Categorization Accuracy for Four Models Across Two Features(Mental Health Datasets)

Class	Accuracy(%)	Precision(%)	Recall(%)	F1-score(%)
Machine_generated	69.00	69.25	66.75	68.00
Human_generated	68.00	68.00	70.75	69.00

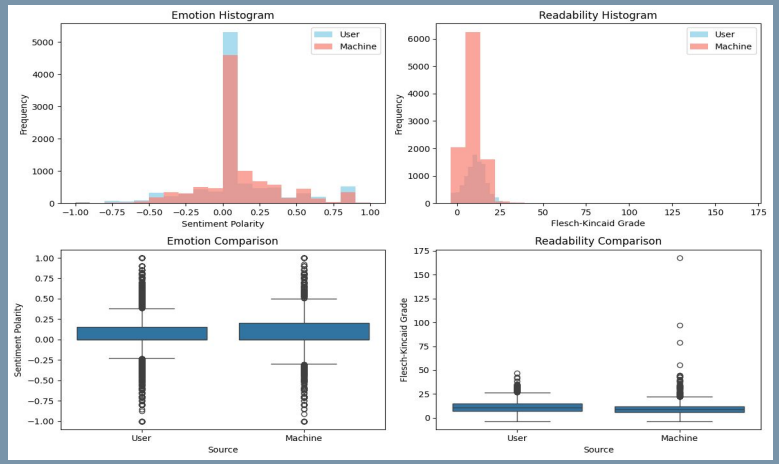
Table 6. Summary of Categorization Accuracy for Four Models Across Two Features(US Airlines Datasets)

Sentiment, Readability and Personality Analysis

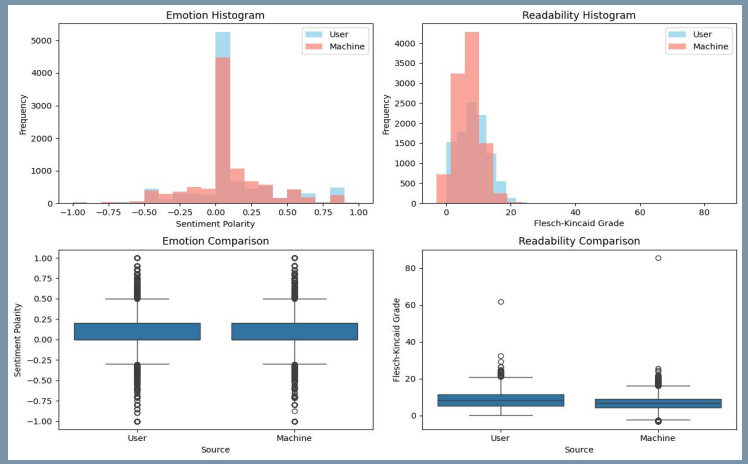
- Conducted a comparative analysis of semantic and syntactic variations between user-generated and machine-generated tweets across 5 domains: FIFA, Game of Thrones and election, utilizing advanced NLP techniques.
- Curated two distinct datasets for each domain, comprising user-authored tweets and machine learning model-generated content.
- Performed sentiment analysis using TextBlob to measure emotional valence and employed textstat to evaluate readability via the Flesch-Kincaid grade level.
- Applied independent sample t-tests to determine the statistical significance of the differences in sentiment and readability between the two sets of tweets.
- Ensured methodological consistency by replicating the analysis process across all domains for a comprehensive comparative evaluation.
- Tweets were tokenized using BertTokenizer and processed through BertForSequenceClassification to predict personality traits, leveraging a GPU environment for efficiency.
- The model's output logits were converted into a probability distribution to score and identify the likelihood of each personality trait, such as Extraversion and Neuroticism, for interpretation.

Sentiment and Readability Analysis

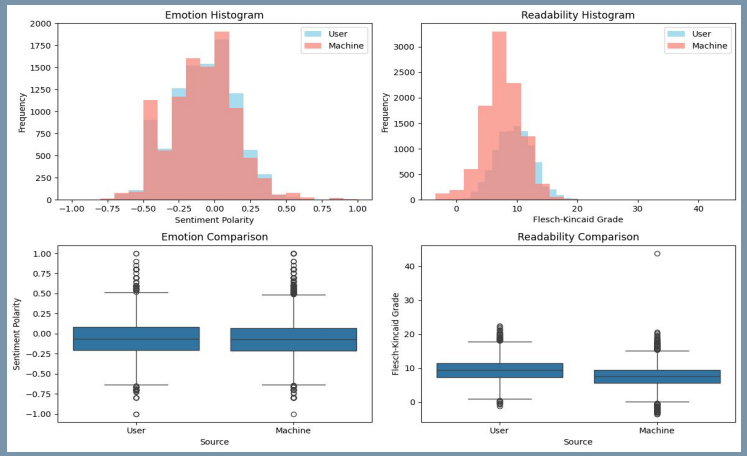
Election



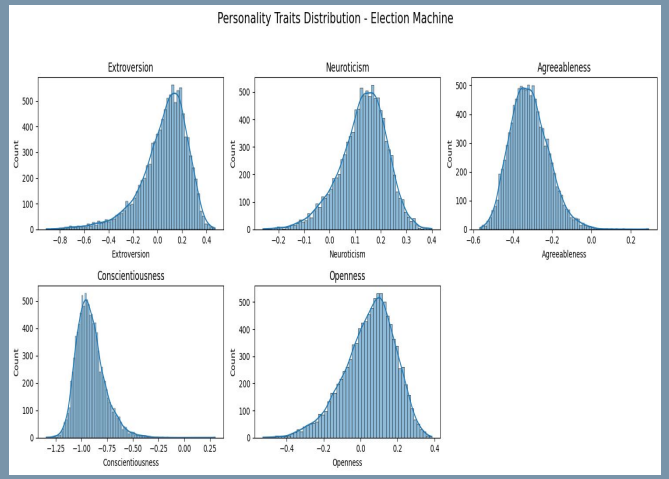
Fifa

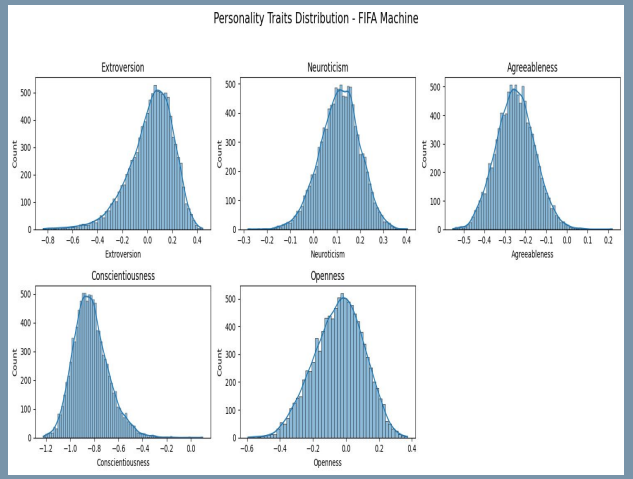


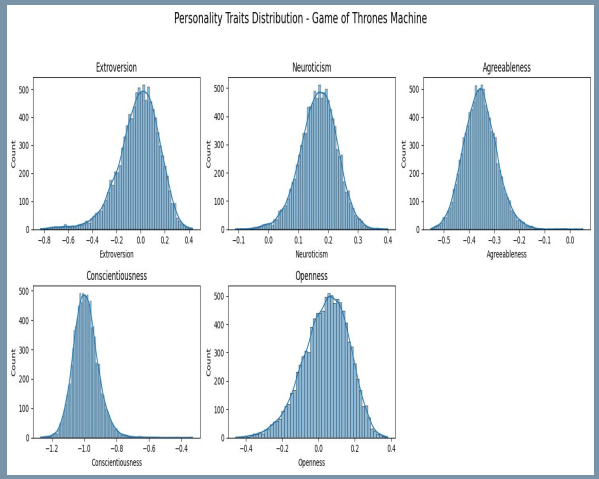
Game of Thrones

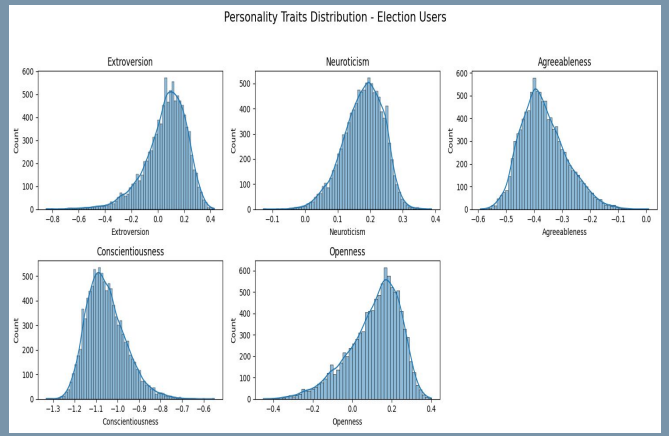


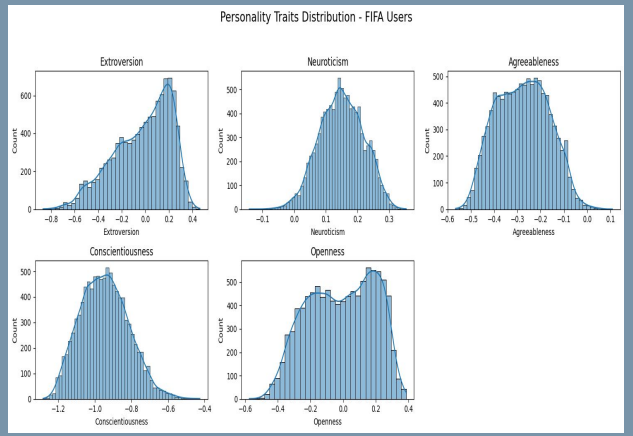
Personality Analysis

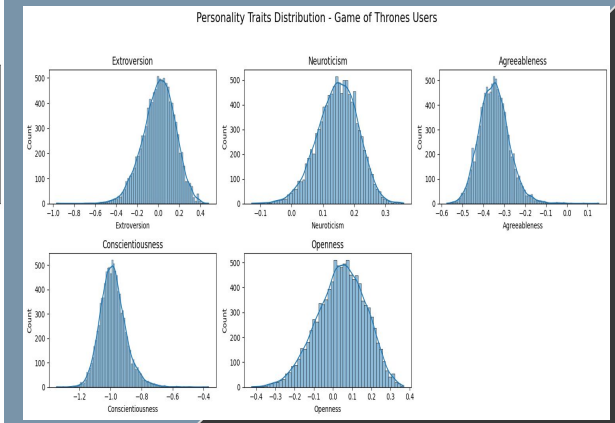












Overall Analysis

- User-generated tweets exhibit more variability in sentiment and readability, reflecting a wider range of emotions and more complex language use.
- Machine-generated content tends to have a neutral tone with less variation in sentiment and a more uniform readability level.
- The simplicity of language in machine-generated tweets compared to user-generated ones suggests room for improving the sophistication of language models.
- Machine-generated data for the Election, FIFA, and Game of Thrones datasets show nearly perfect normal distributions, indicating standardized or normalized data generation processes.
- User-generated data exhibits more irregularities and variability, with traits like 'Conscientiousness' and 'Openness' showing notable skewness, reflecting more natural data variations.
- 'Neuroticism' displays the most consistent distribution across all user and machine datasets, while 'Conscientiousness' and 'Openness' vary more significantly, hinting at their sensitivity to the context of the dataset.

Future Work

1) Explain difference between results obtained from the three topics:
   - Game of Thrones dataset has significantly higher accuracy
2) Optimize existing model: Employ deep learning techniques
3) Pool data from all topics and perform an overall assessment

THANK YOU

Any Questions, feel free to ask !