Data Visualization Project

Bot Detection Dashboard for Social Media Accounts

Fahim Shahriar
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
fasha23@student.sdu.dk

Adib Abzaal
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
adabz23@student.sdu.dk

December 14, 2025

Percent of Contribution (PoC)

Percent of Contribution for the Group Project

Student	**SDU ID / E-mail**	**Percent of Contribution (%)**
Fahim Shahriar	`fasha23@student.sdu.dk`	100
Adib Abzaal	`adabz23@student.sdu.dk`	100

Abstract

1 Background and Motivation

2 Project Objectives

List the research questions:

1. **RQ1:** How do bots differ from humans in follower–following behaviour, profile completeness, and user rank?

2. **RQ2:** Do bots and humans differ in posting frequency and the engagement they receive (comments, reposts, likes)?

3. **RQ3:** Are there differences in URL usage, text length, punctuation, and emotion usage between bots and humans?

4. **RQ4:** Are there differences in posting time behaviour (average posting time, variability, and time-related parameters) between bots and humans?

3 Data

3.1 Data Source

3.2 Data Description

3.3 Data Processing

Before creating the visualizations, several data processing steps were applied to ensure that the dataset was suitable for visual analysis and comparison between bot and human accounts. The original dataset (`SocialBot.xlsx`) contains account-level features related to profile characteristics, activity patterns, engagement metrics, and content properties.

First, categorical variables were cleaned and standardized. The variable `is_bot`, originally encoded as binary values, was converted into a categorical factor with meaningful labels (*Human* and *Bot*) to improve readability in the visualizations. In addition, user rank values (`urank`) were discretized into three rank groups (*Low*, *Medium*, and *High*) to support grouped comparisons and animated storytelling.

Missing values were handled conservatively to preserve as much data as possible. For aggregated statistics such as mean engagement values (likes, reposts, and comments), missing observations were ignored using pairwise deletion. For multivariate analyses such as correlation matrices and principal component analysis (PCA), only complete numeric records were used to ensure numerical stability and avoid biased results.

Several derived features were created to better capture behavioural differences between bots and humans. These include follower–following ratios, posting rate summaries, engagement averages, URL usage statistics, text length variability, and emotion-related content features. For visualization purposes, selected variables were reshaped into long format, enabling the use of faceted plots and grouped bar charts.

Lastly, to make sure that variables with different scales contributed equally to the results, numerical features were normalized where necessary, especially for PCA and clustering analyses. Maintaining transparency, reducing needless data loss, and preparing the dataset to enable understandable visual comparisons between bot and human accounts were the main goals of the data processing.

4 Visualization and Dashboard

4.1 Design Choices

The dashboard was designed with the goal of supporting visual analysis while remaining user friendly design. To achieve this, the interface follows a clear and structured layout where different analytical perspectives are separated into tabs with icons. Each tab corresponds to a specific research question or analysis goal, allowing users to gradually move from an overview of the dataset to more detailed and interactive explorations.

A sidebar-based navigation layout was chosen to provide consistent and intuitive access to all dashboard components. The main tabs include *Overview*, *Profile & Popularity*, *Activity & Engagement*, *Content & Timing*, *Principal Component Analysis (PCA)*, *Interactive Story*, and *AI-generated Graph*. This organisation helps users maintain context and reduces cognitive load by grouping related visualizations together.

Visual encoding choices were made to ensure clarity and consistency across the dashboard. Account type is encoded using colour, with human accounts shown in green or blue and bot accounts shown in red, making comparisons immediately visible across all plots. Quantitative variables such as follower–following rate, posting rate, engagement metrics, and content features are mapped to position and scale, which are known to be among the most accurate visual encodings for numerical data. Where appropriate, faceting is used to display multiple related measures within a single figure without overcrowding the visual space.

Interactivity plays a central role in the dashboard design. Users can explore patterns through hover tooltips, dynamic filters, and animated transitions. In particular, the interactive story tab allows users to examine how engagement behaviour changes under different follower–following rate scenarios, while the decision boundary and clustering tools support deeper analytical exploration. Animation is used selectively to highlight behavioural changes over stages rather than as a decorative element.

4.2 Features

The dashboard was designed and implemented to fully satisfy all mandatory requirements specified in the course project guidelines. Each must-have feature was integrated in a way that supports both analytical exploration and clear communication of results.

At first, the dashboard includes a wide range of visualization types in order to capture different aspects of the data. These include bar charts, boxplots, scatter plots, histograms, heatmaps, and parallel coordinate plots. For instance, boxplots are used to compare follower–following ratios and user rank between human and bot accounts (see figs. 1 and 2), while histograms provide insight into the distribution of follower–following rates across account types (see fig. 3). Scatter plots and PCA-based visualizations further support multivariate analysis (see figs. 6 and 8).

Second, at least one animated visualization is included to illustrate how behavioural patterns change under different scenarios. The animated story focuses on engagement behaviour across varying follower–following rate levels, allowing users to observe how average likes shift dynamically rather than relying on static comparisons. A representative frame of this animation is included in the appendix (see fig. 14).

Third, an AI-generated visualization is incorporated to provide a high-level summary of relationships between numeric features. Specifically, an AI-assisted correlation heatmap is used to highlight global associations among behavioural, engagement, and content-related variables. This visualization supports exploratory reasoning by revealing patterns that may not be immediately visible in individual plots (see fig. 9).

So, the dashboard contains well over the required nine visualizations, distributed across thematic tabs such as Overview, Profile & Popularity, Activity & Engagement, Content & Timing, PCA, and the Interactive Story. Several interactive analytical tools are also provided, including a clustering explorer, a decision boundary visualisation, and a bot detection sandbox, which allow users to experiment with feature values and observe model-driven responses (see figs. 11 to 13). Additionally, the dashboard has the ability to export documentation and results.

4.3 Special Features

The Interactive Story, an important optional feature of the dashboard, allows actively investigate behavioral patterns and model behavior, going beyond conventional static visualizations. This section uses model-based visual analytics, animation, and interaction to facilitate exploratory learning and interpretation. The tab contains following story tabs:

- The *Animated Story* tab presents an animated scatter plot that illustrates how engagement changes under different follower–following rate scenarios. By stepping through predefined stages, observe how average likes vary across account types and rank groups, making behavioural differences easier to interpret over changing conditions (see fig. 14).
- The *Decision Boundary* tab visualises the output of a logistic regression model by displaying predicted bot probabilities as a background surface, with actual accounts overlaid as points. This helps understand how the model separates bots from humans across different feature combinations and supports transparency in model-driven analysis (see fig. 13).
- The *Clustering Explorer* tab enables unsupervised exploration using k-means clustering. By adjusting the number of clusters and selected features to observe how accounts group together and how bot and human accounts are distributed across clusters (see fig. 12).
- The *Detection Sandbox* provides a parallel coordinates view in which define a hypothetical account by adjusting feature values. The resulting scenario is visualised alongside real accounts, allows intuitively compare multivariate behaviour and examine how different feature settings relate to bot detection (see fig. 11).
- The *Feature Importance* tab displays the relative importance of features derived from a logistic regression model. This view supports interpretability by showing which variables contribute most strongly to distinguishing bot accounts from human accounts (see fig. 10).

4.4 Dashboard Link and Report Download

The interactive dashboard developed for this project is publicly accessible online and allows users to explore all visualizations, animations, and interactive analysis components in real time. A permanent link to the deployed dashboard is provided in the references section of this report (see [1]).

In addition to the live dashboard, a report download feature is integrated directly into the interface to support documentation and offline review.

The project contains two download functionality:

1. *Analysis report* Automatic generation of a PDF report that summarises the dataset, visualizations, and key analytical insights.
2. *Project report* manually download the project report from the dashboard where project visualization analysis are briefly described.

5 Story and Results

This section presents the main findings of the project by answering the research questions introduced earlier (see **RQ1**, **RQ2**, **RQ3**, **RQ4**). The results are derived from the visual patterns observed across the dashboard and are supported by both static and interactive visualizations. Where appropriate, additional figures included in the appendix are referenced to provide further evidence.

5.1 RQ1 – Profile and Popularity

Research Question 1 investigates how bot accounts differ from human accounts in terms of follower–following behaviour, profile-related indicators, and user rank. The visualizations reveal clear and consistent differences between the two account types.

Boxplots of the follower–following ratio show that bot accounts tend to exhibit more extreme values than human accounts, with a wider spread and a higher number of outliers (see fig. 1). This suggests that bots often follow disproportionate strategies, either following many accounts with little reciprocity or maintaining unusually high follower ratios. In contrast, human accounts generally display more moderate and balanced follower–following behaviour.

A similar pattern is observed in the distribution of user rank. As shown in fig. 2, bot accounts are more frequently associated with lower rank values, while human accounts tend to occupy higher and more stable rank ranges. These findings indicate that profile-level popularity metrics can provide useful signals for distinguishing between bots and humans.

5.2 RQ2 – Activity and Engagement

Research Question 2 focuses on differences in posting behaviour and engagement metrics between bot and human accounts. The dashboard visualizations reveal meaningful contrasts in both activity levels and user responses.

The distribution of posting rates indicates that bot accounts often post more frequently than human accounts, with a broader and more skewed distribution. This behaviour reflects automated or semi-automated posting strategies, whereas human accounts typically show more moderate posting patterns. Engagement-related visualizations further highlight these differences. As illustrated in the engagement comparison plots, human accounts tend to receive higher average levels of likes, comments, and reposts, while bot accounts generally experience lower engagement despite higher activity levels.

The histogram of follower–following rates by account type (see fig. 3) also supports this interpretation, as bots with extreme activity patterns do not necessarily translate their activity into meaningful engagement. This suggests that engagement metrics provide complementary information to activity-based features when analysing account behaviour.

5.3 RQ3 – Content Style

5.4 RQ4 – Temporal Patterns

6 Conclusion and Discussion

References

[1] F. Shahriar and A. Abzaal. *SocialBot Bot Detection Dashboard*. Available at: https://019b1f12-b9c4-4abe-b67a-8df144c084fd.share.connect.posit.cloud/

A Appendix: Additional Visualizations

A.1 Profile & Popularity (RQ1)

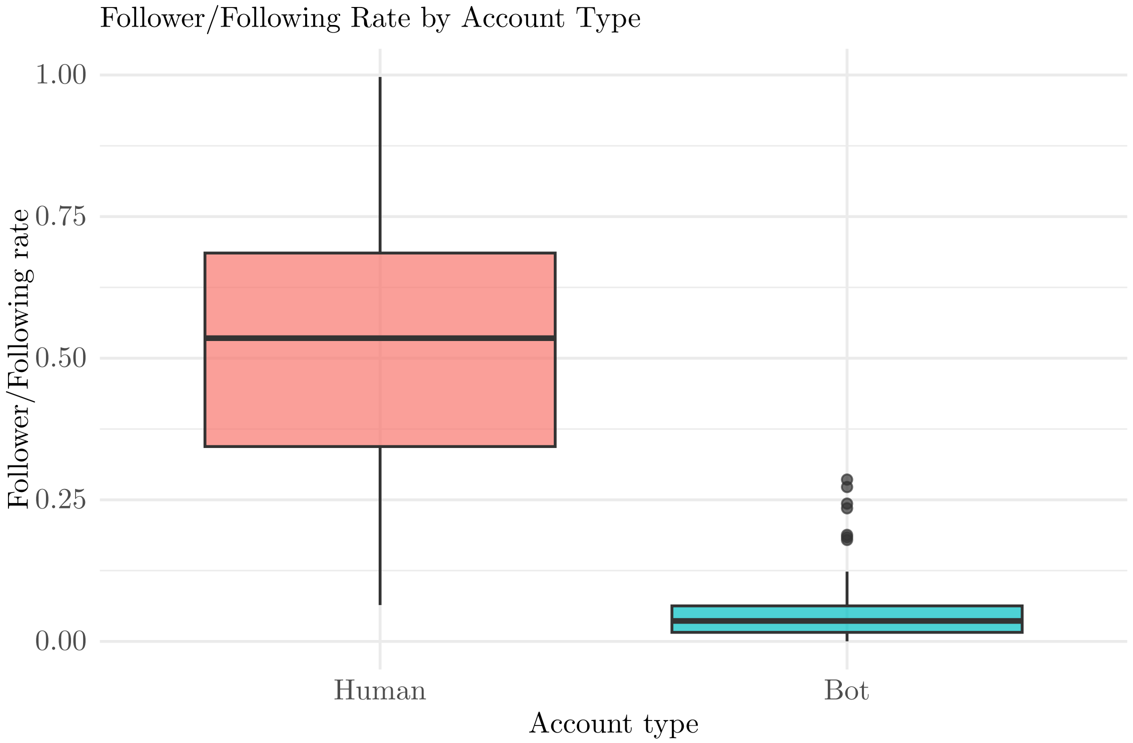


Figure 1: Follower–Following ratio comparison between Human and Bot accounts.

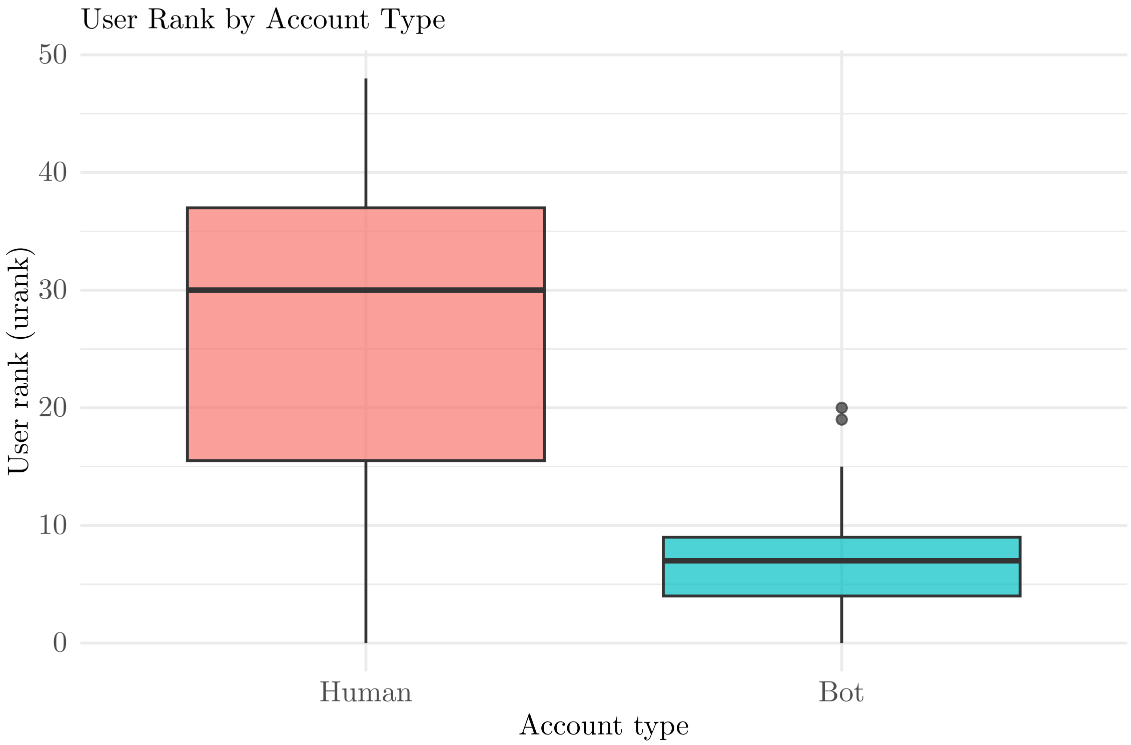


Figure 2: Distribution of user rank (urank) for Human and Bot accounts.

A.2 Activity & Engagement (RQ2)

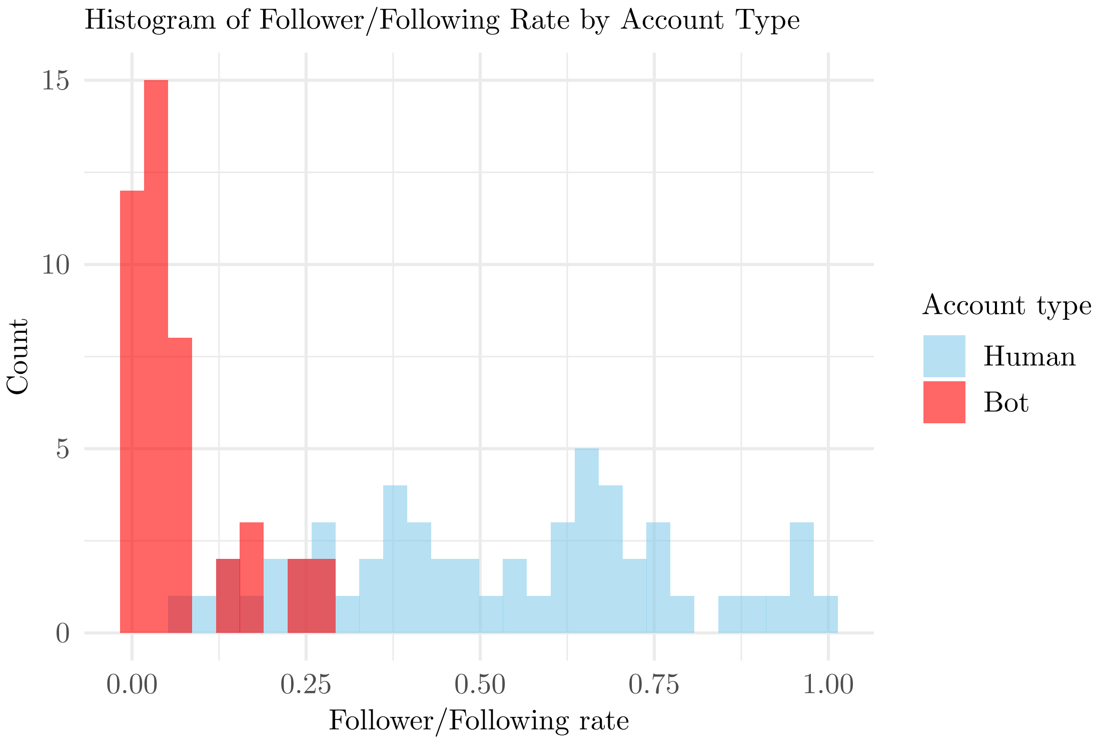


Figure 3: Histogram of follower–following rate by account type.

A.3 Content Style & Timing (RQ3–RQ4)

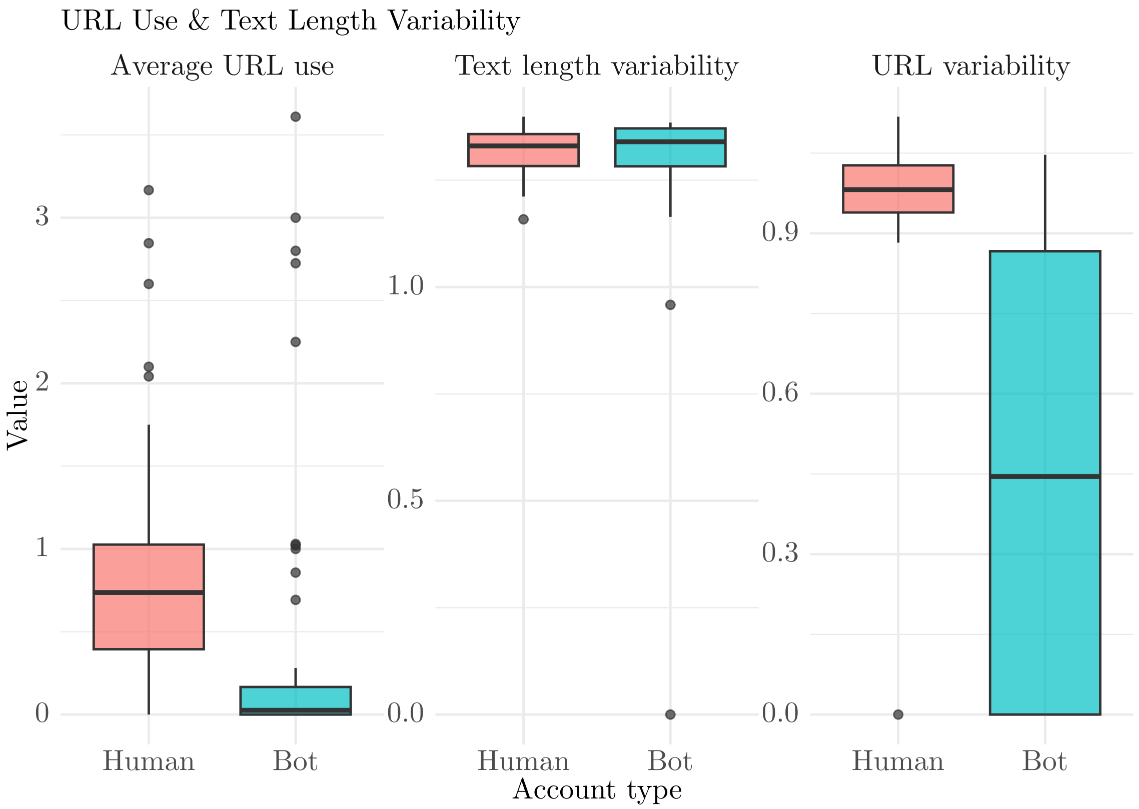


Figure 4: URL usage and text length variability across Human and Bot accounts.

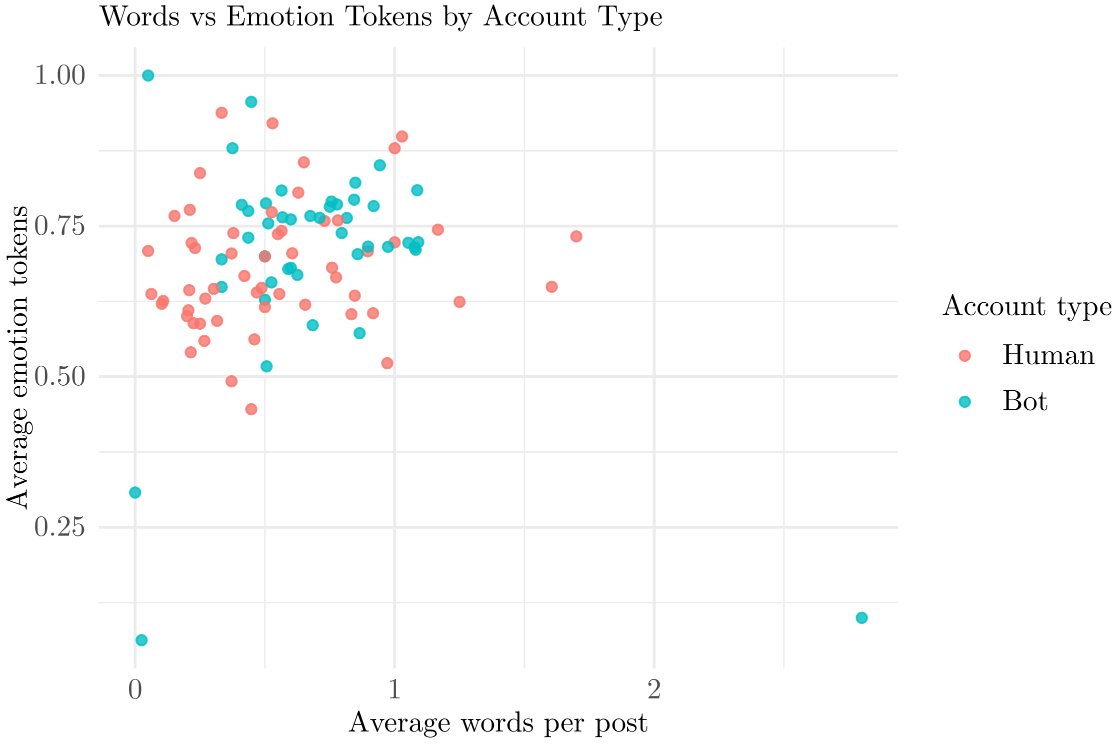


Figure 5: Relationship between average words per post and emotion tokens.

A.4 PCA Visualizations

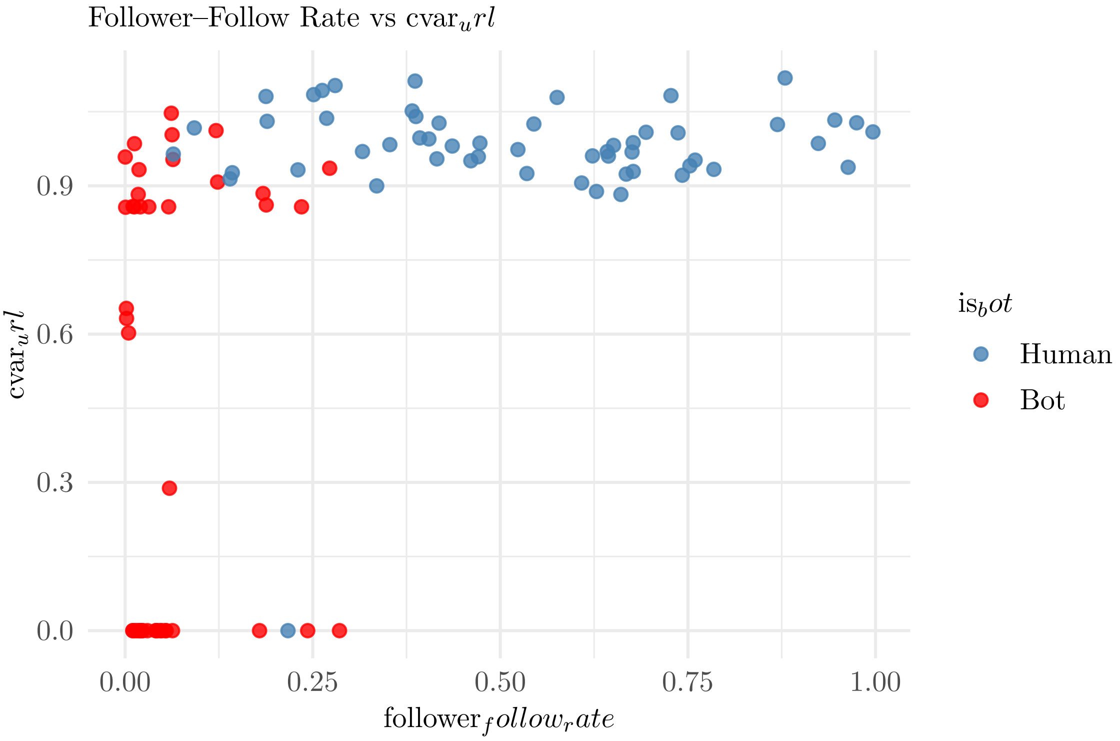


Figure 6: Pairwise scatter plot of follower–following rate vs URL variability (cvar_url).

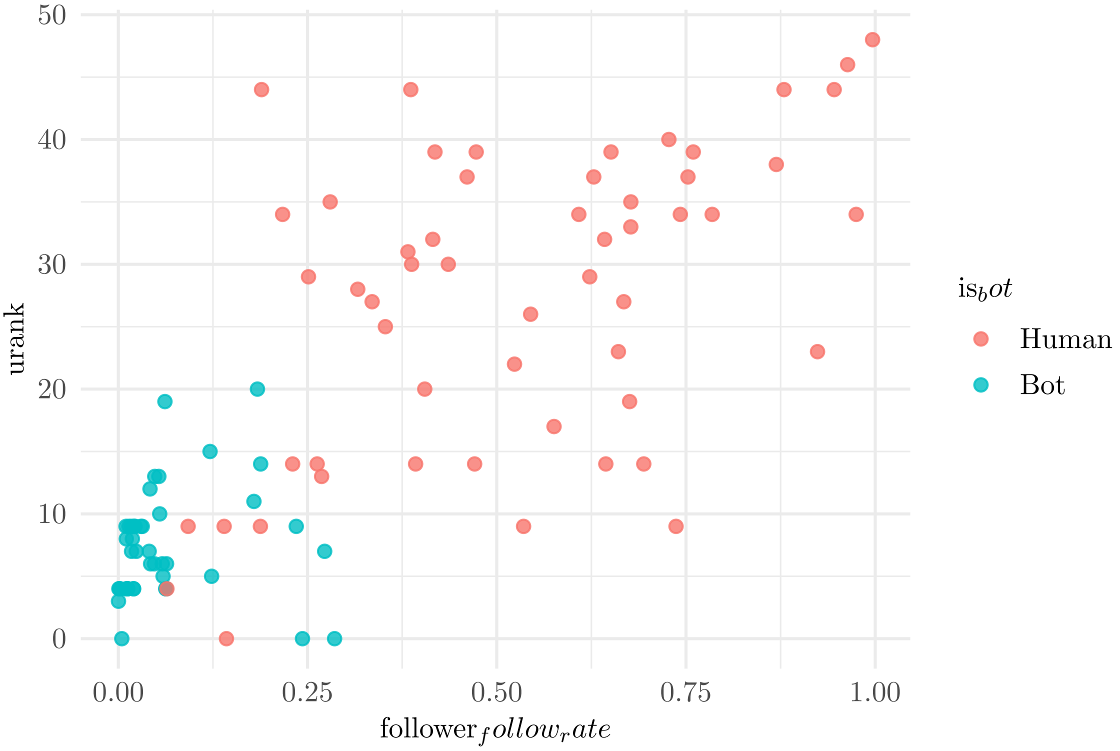


Figure 7: Pairwise scatter plot of follower–following rate vs user rank (urank).

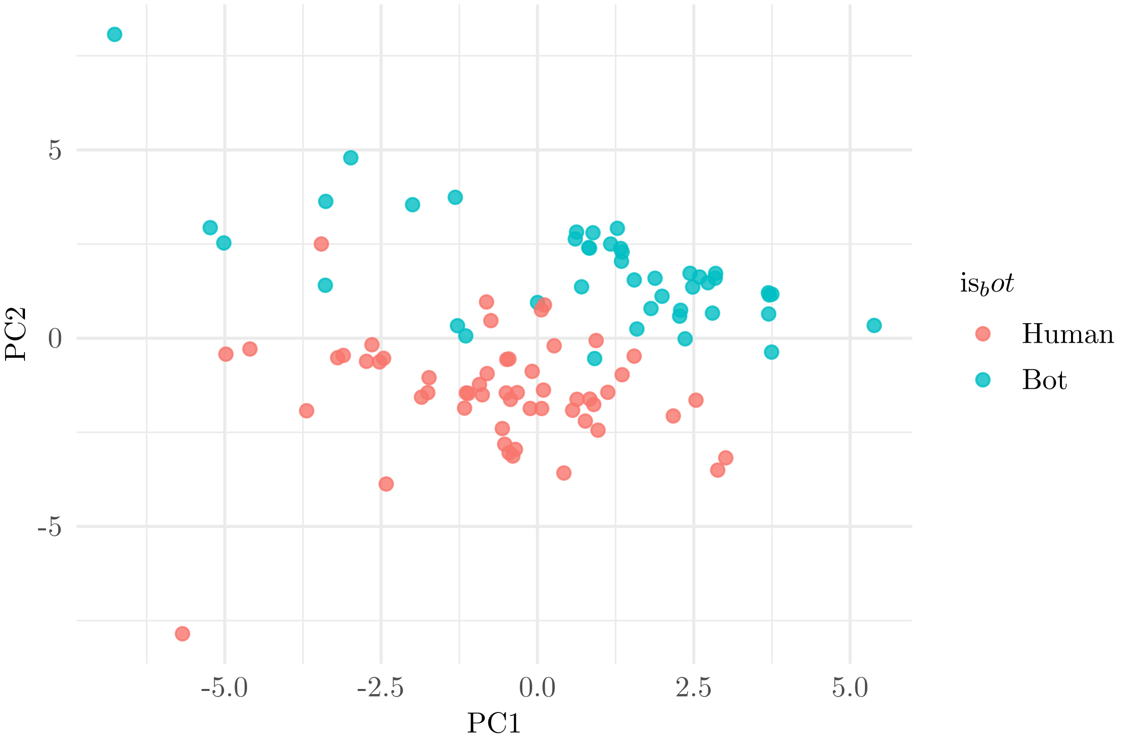


Figure 8: Principal Component Analysis (PC1 vs PC2) scatter plot coloured by account type.

A.5 AI-generated Correlation Heatmap

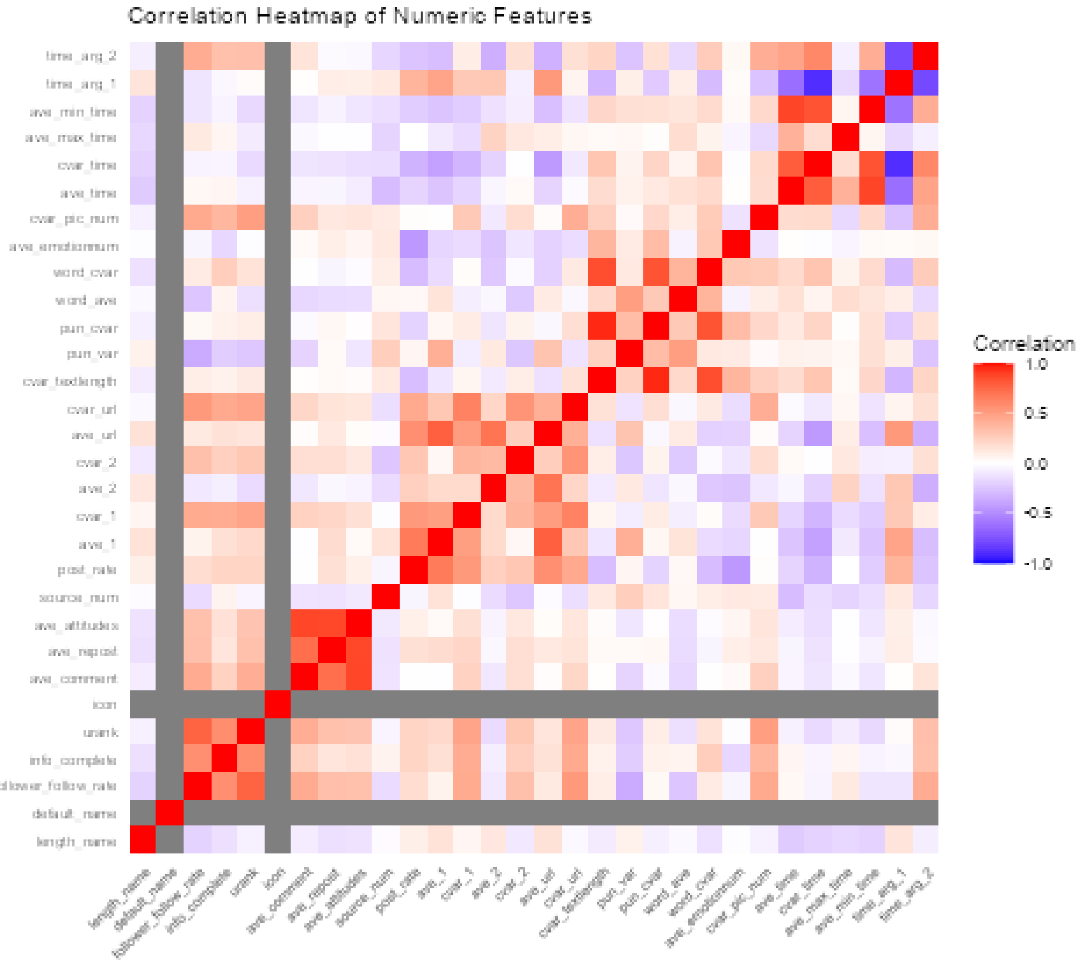
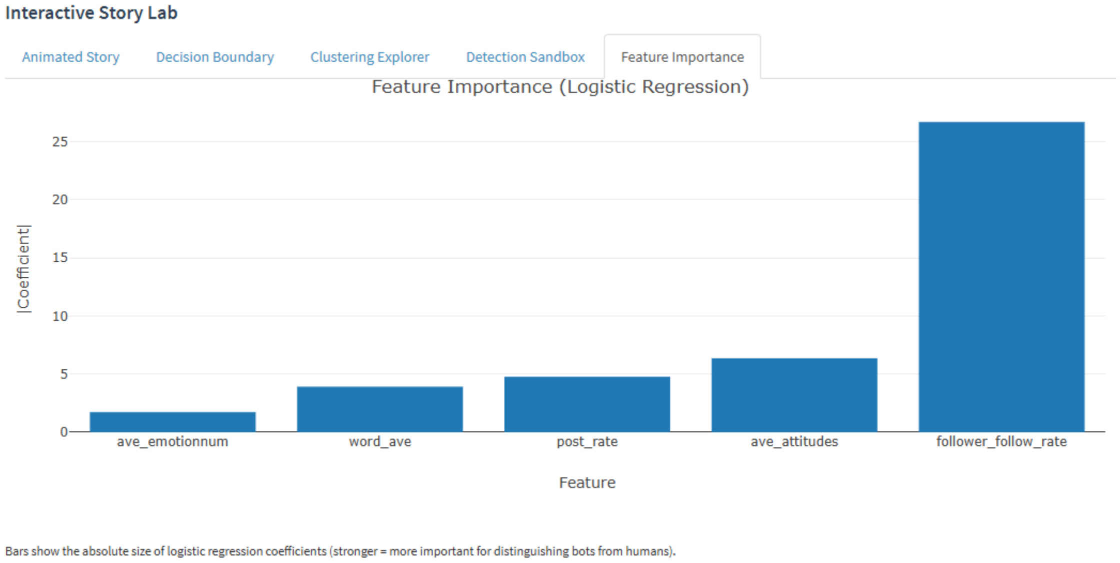


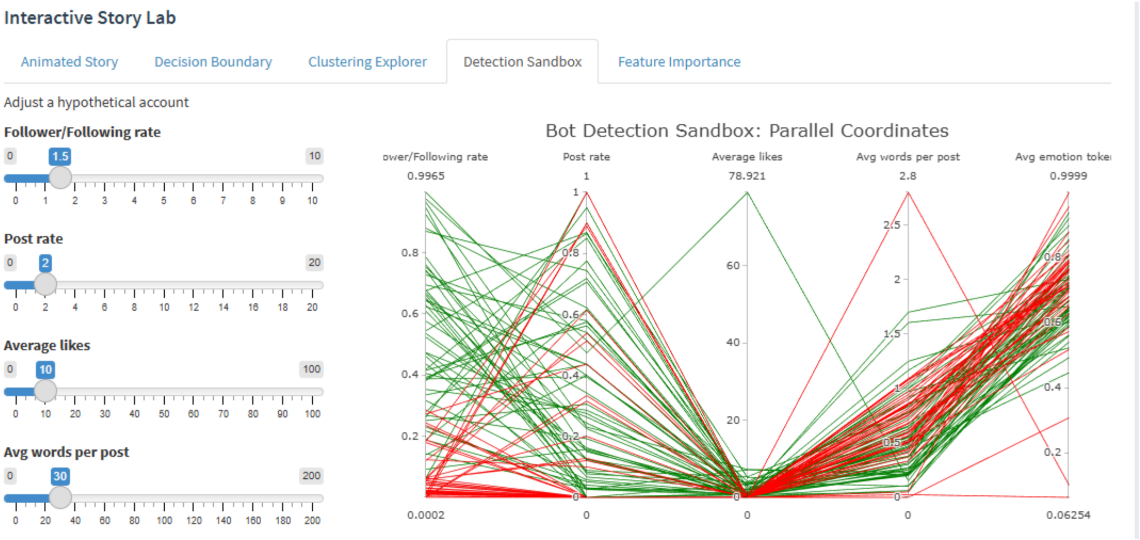
Figure 9: AI-generated correlation heatmap of numeric features, summarising global relationships between behavioural and content variables.

A.6 Interactive Story: Additional Screenshots



Bars show the absolute size of logistic regression coefficients (stronger = more important for distinguishing bots from humans).

Figure 10: Feature importance derived from logistic regression. Bars represent the absolute magnitude of model coefficients, indicating the relative contribution of each feature to distinguishing bot and human accounts.



Adjust a hypothetical account

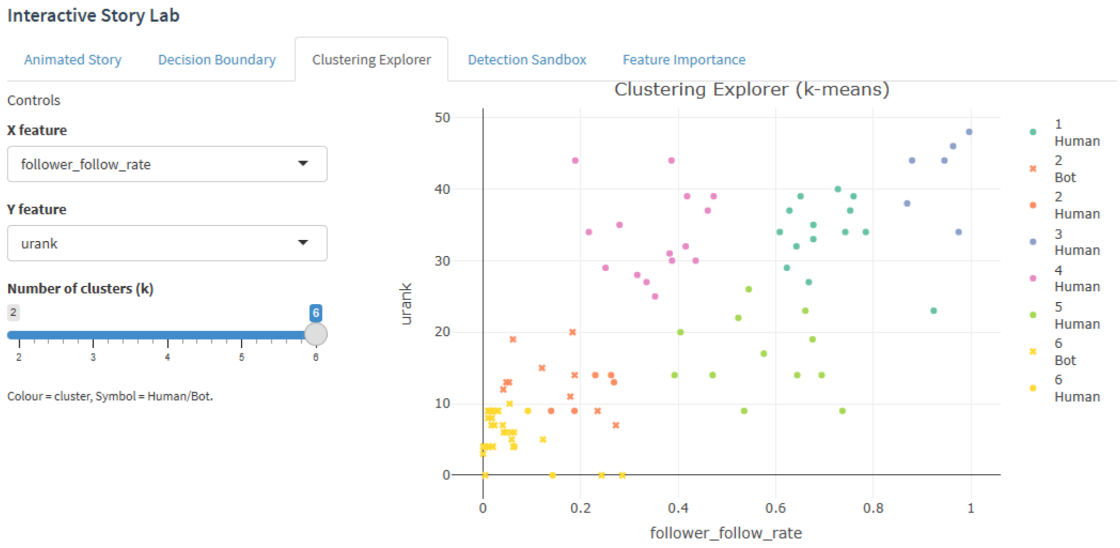
Follower/Following rate: 1.5 (0–10)

Post rate: 2 (0–20)

Average likes: 10 (0–100)

Avg words per post: 30 (0–200)

Figure 11: Bot Detection Sandbox using parallel coordinates. Each line represents an account, with colour indicating account type (Human or Bot). The highlighted scenario line shows a user-defined hypothetical account.



Interactive Story Lab

Controls

X feature

follower_follow_rate

Y feature

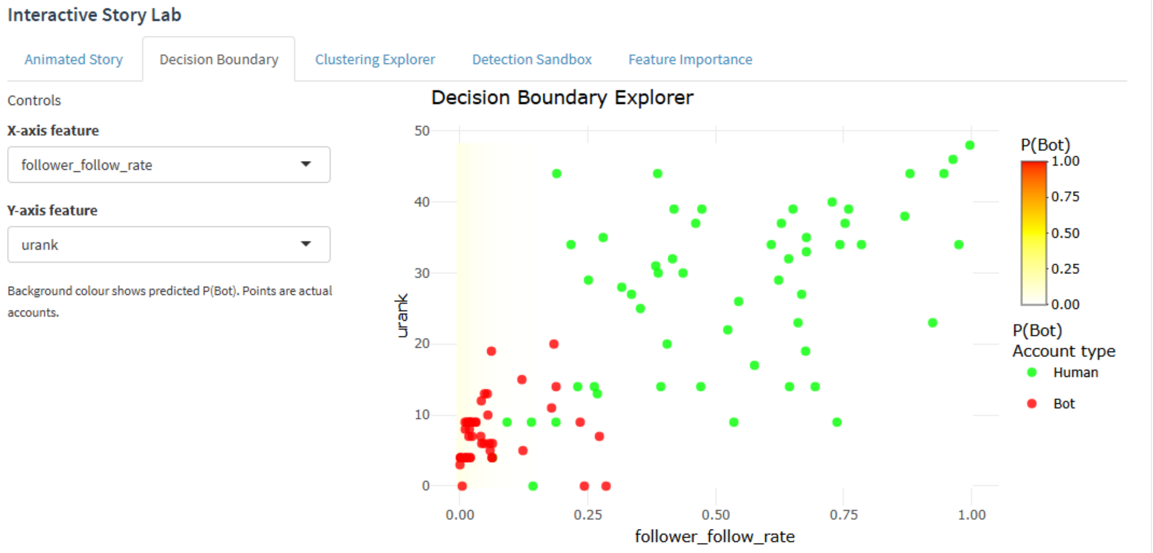
urank

Number of clusters (k)

2 — 6

Colour = cluster, Symbol = Human/Bot.

Figure 12: Clustering explorer based on k-means clustering. Points are coloured by cluster membership, while symbol shape indicates account type.



Interactive Story Lab

Controls

X-axis feature

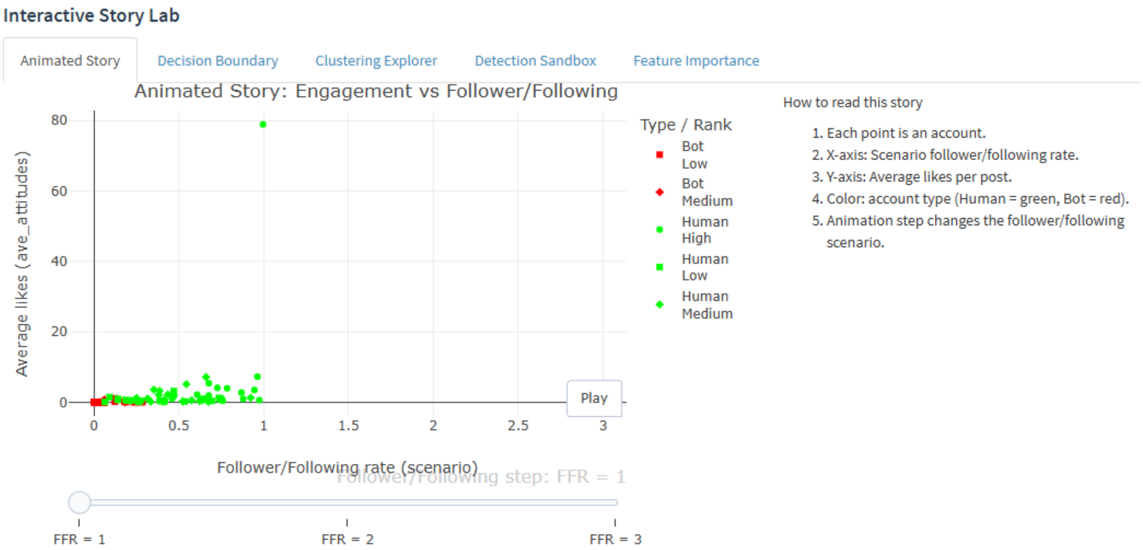
follower_follow_rate

Y-axis feature

urank

Background colour shows predicted P(Bot). Points are actual accounts.

Figure 13: Decision boundary explorer based on a logistic regression model. Background colour represents predicted bot probability, while points show actual accounts.



How to read this story

1. Each point is an account.
2. X-axis: Scenario follower/following rate.
3. Y-axis: Average likes per post.
4. Color: account type (Human = green, Bot = red).
5. Animation step changes the follower/following scenario.

Figure 14: Animated story visualising engagement behaviour under different follower–following rate scenarios. Animation highlights how average likes vary across account types and rank groups.