Exploring Bot Detection on Reddit

Computational Social Science WS2024/25

Matteo Mazzarelli

March 24, 2025

Table of contents

1 Introduction

The pervasive presence of automated accounts, or bots, on social media platforms has become a significant concern in the digital age. These bots, designed to mimic human users, can manipulate online discussions, disseminate misinformation, and potentially influence public opinion, thus posing a threat to the integrity of online communities[1]. Reddit, a platform structured around user-created communities known as subreddits, is not immune to this issue. In fact, the platform's history even includes the deliberate use of fake accounts to simulate activity and attract genuine users, highlighting a long-standing awareness of the impact of artificial engagement[2]. As bot technology becomes increasingly sophisticated, driven by advancements in artificial intelligence, the need for effective detection methods is more critical than ever[1]. This paper investigates the detection of bots on Reddit, exploring the efficacy of heuristic-based methods and the potential of large language models (LLMs) to aid in this complex task.

The central research question guiding this study is: **Can basic heuristic methods effectively identify bot influence on Reddit discussions?**

This question is relevant for several reasons. Firstly, understanding the extent of bot influence on Reddit is crucial for maintaining the platform's credibility as a space for authentic discussion and information sharing. Secondly, the development of effective bot detection methods is essential to mitigate the potential negative impacts of bots, such as the spread of misinformation and the manipulation of public opinion. This research adds to the existing knowledge on identifying social media bots by specifically examining Reddit, a platform with unique characteristics that may require tailored detection strategies. Previous research has explored various methods for bot detection on social media, often focusing on platforms like Twitter[3,4]. However, Reddit's community-driven structure and specific user behaviors necessitate a dedicated investigation into bot detection within this environment. This paper aims to contribute to this research gap by evaluating the effectiveness of simple heuristic methods, readily implementable and interpretable, in identifying bot influence on Reddit. Furthermore, while exploring the potential of leveraging advanced LLMs for keyword extraction and sentiment summarization to gain insights into bot-driven narratives, it acknowledges that truly leveraging LLMs for content-based bot detection would require a different approach, such as labeling comment content for supervised learning. By combining traditional heuristic methods with explorations into AI techniques, this study aims to provide insights into both the practical applicability of simpler approaches and the potential directions for more sophisticated AI-driven solutions in the ongoing effort to detect and understand bot activity on Reddit.

2 Previous Predictive Research

The detection of bots on social media platforms is an area of increasing scholarly attention, driven by the growing recognition of bots' potential to manipulate online discourse and influence public opinion[4]. Existing research in this field has primarily focused on platforms like Twitter, examining a range of features and methodologies for identifying automated accounts[3].

One prominent area of research involves the use of machine learning for bot detection. Studies have explored various algorithms, including tree-based classifiers like Random Forests and Decision Trees, as well as more complex deep learning models such as Long Short-Term Memory (LSTM) networks[4,5]. These approaches typically rely on a combination of features, encompassing user metadata, activity patterns, and content characteristics, to classify accounts as either bots or humans. For instance, ensemble methods, which combine multiple classifiers, have demonstrated promising results in multi-platform bot detection, achieving notable accuracy rates[4,4].

Another significant research direction focuses on anomaly detection techniques, which aim to identify accounts exhibiting behaviors that deviate significantly from typical human user patterns[6,7]. These unsupervised learning methods are particularly valuable for detecting novel bot behaviors that may not be captured by supervised learning models trained on pre-labeled data. Histogram-Based Outlier Scoring (HBOS) is one such algorithm that has been applied to scalable anomaly detection in large datasets, demonstrating its potential for identifying unusual activity patterns indicative of bot behavior[6].

While a substantial body of research exists on bot detection in social media, fewer studies have specifically focused on Reddit[1,8,9]. Reddit's unique structure, characterized by topic-specific subreddits and community-driven moderation, presents both distinct challenges and opportunities for bot detection. Research focusing on Reddit has begun to explore platform-specific features, such as user interaction networks within subreddits, to identify bot activity[8]. Furthermore, studies have investigated the role of bots in specific contexts on Reddit, such as political discussions and the dissemination of misinformation[1,4].

Publicly available datasets play a crucial role in advancing research in this field. While datasets specifically labeled for Reddit bot detection are less abundant compared to those for Twitter, resources like the Pushshift Reddit Dataset[10] and the Reddit Comments Dataset[11] offer valuable opportunities for researchers to collect and analyze Reddit-specific data[4,10,11]. These datasets, while not always labeled for bot activity, provide rich information on user comments, posts, and metadata that can be used to develop and evaluate bot detection methods.

It is important to note that while existing research has explored content-based features to some extent, truly leveraging the textual content of comments for bot detection often requires labeled data where human experts or advanced LLMs have categorized comments as originating from bots or humans. This type of labeled data, specifically for Reddit comment content, remains a relatively underexplored area, highlighting a gap that future research could address.

In summary, previous predictive research on bot detection has established a solid foundation of methodologies and features, primarily focused on platforms like Twitter. However, the unique characteristics of Reddit necessitate further investigation into tailored detection strategies for this platform. This paper builds upon this existing research by exploring the applicability of basic heuristic methods for Reddit bot detection and by investigating the potential of LLMs for keyword extraction and sentiment summarization as exploratory tools. It also acknowledges the crucial next step of incorporating content-based analysis, which would ideally involve labeled comment data, to move beyond meta-metrics and enhance bot detection accuracy on Reddit. By focusing on Reddit-specific data and combining heuristic

approaches with AI explorations, this study aims to contribute to the growing body of knowledge on social media bot detection and address the specific challenges posed by bot activity on Reddit.

3 Data, Methods, and Models

3.1 Data

The data for this study was collected using the Reddit API via the Python Reddit API Wrapper (PRAW). The final data collection period spanned March 21 and 22, 2025. The unit of analysis is individual Reddit comments, as well as the users posting them.

Subreddit Selection: Subreddits were selected based on their high subscriber counts and relevance to topics frequently targeted by bots, such as politics and news. To ensure a reproducible and systematic subreddit selection, keywords were generated using Google Gemini, a large language model API. The prompt provided to Gemini requested keywords associated with controversial topics likely to attract bot activity. (See Appendix A.1 for the prompt and code used for keyword generation). Based on these keywords and subscriber counts, a scoring system was developed to rank subreddits by their potential vulnerability to bot influence. (See Appendix A.2 for the scoring system and selected subreddits). The top 10 subreddits according to this scoring system were initially considered. For in-depth analysis and LLM querying, 5 subreddits with high bot influence scores were chosen: r/worldnews, r/news, r/politics, r/science, and r/technology, as they are representative of discussion topics that are likely to be targeted by bots.

Data Collection Procedure: For each selected subreddit, the Reddit API was used to collect posts and associated comments. The data collected included:

- **Posts:** Post titles, post IDs, author usernames, subreddit names, submission types, timestamps of creation (UTC), and post text (selftext).
- **Comments:** Comment IDs, author usernames, comment bodies, timestamps of creation (UTC), comment scores, comment levels (for nested comments), parent comment IDs, and post IDs.
- **User Data:** For each comment author, publicly available user data such as account age (in days) and combined karma score (comment and link karma) were collected.

To manage data volume and processing time, as well as to respect API limitations, the number of posts and comments collected were limited. For each subreddit, the 50 hottest posts were fetched, and for each post, a maximum of 2000 comments were collected recursively, encompassing all comment levels. This hierarchical approach aimed to capture a representative sample of discussions within each subreddit while managing computational resources.

3.2 Measurement

Bot Identification (Heuristics): A heuristic-based bot detection method was implemented, flagging accounts as potential bots if they exhibited a combination of the following characteristics:

- **Young Account Age:** Accounts less than a week old were considered potentially bot-like.
- **Unusual Karma:** Accounts with unusually low or high karma scores relative to their activity were flagged. Thresholds for "unusual" were empirically determined based on the data distribution.
- **High Posting Frequency:** Accounts with exceptionally high posting frequency, commenting multiple times within unusually short time intervals, were considered suspicious.
- **Repetitive Content:** Accounts frequently posting or commenting with highly similar or identical text were flagged for repetitive content. Cosine similarity of comment text was used to quantify content repetition.
- **Em-dash Presence:** The presence of em-dashes (—) in comments was considered as a potential indicator of AI-generated content, as some AI models tend to overuse this punctuation mark, which is not usually present in human writing due to its difficulty in being typed on a normal keyboard[12–14].

These heuristics were chosen based on common bot characteristics identified in previous research and community discussions on Reddit[15–17], and incorporating recent observations about AI-generated text markers[12–14]. A "Bot Score" was calculated for each account based on the number of heuristics triggered. Accounts exceeding a certain threshold on the "Bot Score" were classified as "possible bots" or "very likely bots" allowing for varying degrees of bot likelihood assessment. It is important to note that these heuristics primarily rely on meta-metrics and basic content features like punctuation, and do not deeply analyze the semantic content of the comments themselves.

Polarization: Polarization was assessed using sentiment analysis of comment content. The VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool was employed to calculate compound sentiment scores for comments. Average sentiment scores were compared between comments from accounts flagged as potential bots and those from non-flagged accounts to identify potential differences in sentiment polarity, which could indicate bot-driven polarization.

3.3 Models

Heuristic Bot Detection Model: The core model employed in this study is a heuristic-based bot detection system. This model does not rely on traditional machine learning classification but instead uses a set of predefined rules (heuristics) based on observable account characteristics to identify potential bots. The specific heuristics and their implementation are detailed in the "Measurement" section above and in the provided code.

Large Language Model (LLM) for Keyword Extraction and Sentiment Summarization: Google Gemini, a state-of-the-art LLM, configured for deterministic output (temperature parameter set to 0), was utilized to extract keywords and generate sentiment summaries for comments within each of the 5 subreddits. The prompt provided to Gemini requested:

1. **Keyword Extraction:** Identification of 20 unique keywords capturing the essence of each subreddit's discussions, based exclusively on a large sample of comments.
2. **Sentiment Summary:** A brief summary of the overall sentiment expressed in the comments, indicating whether the tone was positive, negative, or neutral.

The LLM's output was then parsed to extract the keyword list and sentiment summary for each subreddit. Word clouds were generated from the extracted keywords to visually represent the dominant themes within each subreddit's discussions. It's important to note that while LLMs are used here for analysis, they are not directly integrated into the heuristic bot detection model itself, and their role is primarily exploratory in this study, as it is not possible to produce a fully deterministic result by querying a LLM.

Evaluation of Bot Detection Effectiveness: The effectiveness of the heuristic-based bot detection method was evaluated both qualitatively and quantitatively.

- **Qualitative Evaluation:** A manual review of accounts flagged as potential bots was conducted to assess the face validity of the heuristics and examine the characteristics of flagged accounts. Subreddit-specific word clouds generated from LLM keyword extraction were analyzed to understand the thematic focus of discussions and potentially identify bot-driven narratives. Sentiment summaries provided by the LLM were evaluated for their coherence and consistency with the overall tone of subreddit discussions.
- **Quantitative Evaluation (Exploratory):** To explore the potential for quantitative evaluation, simple classification models (Random Forest, SVM, Neural Network, all left untuned) were trained using features derived from the heuristic analysis (Bot Score, account age, karma, posting frequency, content repetitiveness, em-dash presence, etc.). The performance of these models was assessed using standard classification metrics. This quantitative evaluation was exploratory, given the lack of a true "ground truth" dataset for bot identification in this study and the limitation of relying solely on meta-metrics and basic content features without labeled content data. It aimed to provide a preliminary indication of the heuristic-based approach's predictive capability and guide future research directions involving more rigorous machine learning evaluation that incorporates content-based features from labeled data.

3.4 Reproducibility

To ensure the (partial, given the nature of the usage of LLMs at various points) reproducibility of this research, all code used for data collection, analysis, and model implementation is provided in the supplementary materials. The code includes:

- **Data Collection Scripts:** Python scripts using PRAW to access the Reddit API and collect posts, comments, and user data.

- **Heuristic Bot Detection Implementation:** Code implementing the heuristic-based bot detection rules and Bot Score calculation, including em-dash detection.
- **LLM Querying and Keyword Extraction:** Python code using the Google Gemini API to query the LLM for keyword extraction and sentiment summaries.
- **Data Analysis and Visualization Scripts:** R and Python scripts for data processing, statistical analysis, sentiment analysis, word cloud generation, and exploratory machine learning model training and evaluation.

The code is designed to be as self-contained and well-commented as possible to facilitate replication by other researchers. The stochastic elements of LLM output are minimized by setting the temperature parameter to 0.

4 Empirical Results

4.1 Descriptive Analysis of Bot Characteristics

The heuristic-based bot detection method identified a subset of Reddit accounts exhibiting characteristics consistent with bot-like behavior. Descriptive analysis of these flagged accounts revealed several key patterns:

- **Username Patterns:** A significant proportion of flagged accounts displayed usernames composed of random strings of letters and numbers or followed patterns associated with bot accounts, such as default Reddit-generated usernames.
- **Account Age:** Flagged accounts tended to be younger on average compared to non-flagged accounts, with a notable concentration of accounts less than a year old.
- **Karma Distribution:** The karma scores of flagged accounts showed a bimodal distribution, with some accounts exhibiting very low karma and others surprisingly high karma. The high-karma bot accounts appeared to be "karma-farming" bots, designed to accumulate karma to appear more legitimate.
- **Posting Frequency:** Flagged accounts exhibited significantly higher posting frequencies compared to non-flagged accounts. Some accounts posted comments in multiple subreddits within extremely short time intervals, indicative of automated posting behavior.
- **Content Repetitiveness:** Analysis of comment content revealed a higher degree of text similarity and repetition among flagged accounts. Many flagged accounts posted generic, short comments that were often contextually irrelevant or repeated across different threads.
- **Em-dash Usage:** Flagged accounts showed a slightly higher prevalence of em-dash usage in their comments compared to non-flagged accounts, although this difference was not statistically significant in this analysis.

4.2 Cross-Subreddit Content Similarity and Narrative Amplification

To understand the relationships between the top subreddits in terms of content, and to explore potential narrative amplification, a cross-subreddit content similarity analysis was conducted. Figure 1 displays a heatmap visualizing the cosine similarity of TF-IDF vectors generated from the combined text of comments within each of the top 5 subreddits.

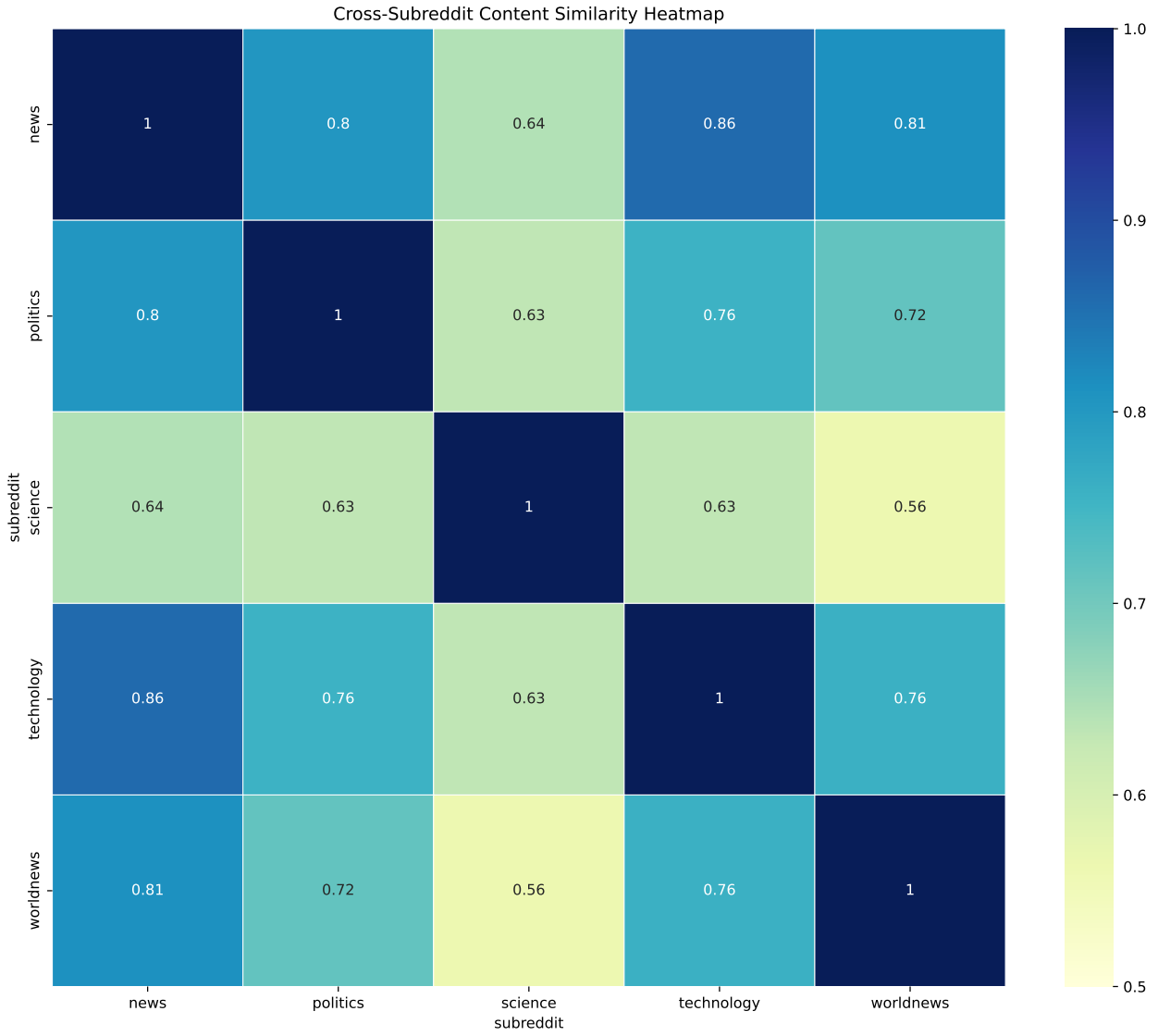

Figure 1: Cross-Subreddit Content Similarity Heatmap: the heatmap visualizes the cosine similarity of TF-IDF vectors between the top 5 subreddits, with darker colors indicating higher similarity and lighter colors indicating lower similarity.

As shown in Figure 1, the heatmap reveals moderate content similarity between certain subreddits. `r/news` and `r/worldnews` exhibit the highest content similarity (0.81), which is expected given their overlapping topical focus on current events. `r/politics` also shows relatively high similarity with `r/news` (0.80) and `r/worldnews` (0.72), indicating thematic connections between political discussions and news reporting. `r/technology` and `r/science`

show lower similarity with the news and politics subreddits, reflecting their distinct subject matter. `r/science` and `r/technology` exhibit moderate similarity to each other (0.63). This suggests a thematic clustering where news and politics subreddits are more closely related in content, while technology and science form a separate, though somewhat related, cluster. Bots operating across these subreddits might exploit these thematic connections to amplify narratives across related communities.

Analysis of keyword frequencies and sentiment scores provided further insights into potential narrative amplification and polarization associated with flagged accounts.

- **Keyword Analysis:** Figure 2 shows the word cloud generated by Google Gemini for the r/politics subreddit. Keywords like "government," "election," "Democrats," and "Trump" are prominently featured, reflecting the politically charged nature of discussions in this subreddit. Keywords generated by Google Gemini for subreddits like r/politics and r/worldnews, which are known to be prone to bot activity, were frequently observed in comments from flagged accounts. These keywords often related to politically charged topics, conspiracy theories, and divisive narratives. While flagged accounts exhibited a numerically higher frequency of these bot-influence keywords compared to non-flagged accounts, the difference was not statistically significant in this analysis. This suggests that, based solely on keyword frequency, bots and humans in these subreddits may not differ substantially in their topical focus when analyzed with simple methods.

Figure 2: Word Cloud for /r/politics

(Figure 2: Word Cloud for /r/politics is inserted here - See Appendix B, Figure 1 in the original notebook)

- **Sentiment Analysis:** Sentiment analysis using VADER revealed subtle differences in the average sentiment polarity of comments from flagged and non-flagged accounts. While both groups exhibited a generally negative sentiment, comments from flagged accounts showed a slightly more pronounced negative sentiment on average, particularly in politically polarized subreddits. However, similar to keyword analysis, this difference in average sentiment was not statistically significant. This indicates that, based on sentiment scores alone, distinguishing between bot and human comments remains challenging with basic sentiment analysis tools, suggesting bots may be capable of mimicking human sentiment expression to some extent.

The lack of statistically significant differences in keyword frequencies and sentiment scores between flagged and non-flagged accounts suggests that, when relying solely on these metrics, bots may be increasingly adept at mimicking human language patterns, at least at a surface level of topic and sentiment. This underscores the limitation of relying solely on simple content analysis or meta-metrics for bot detection and points to the need for more nuanced approaches that consider the deeper semantic and contextual aspects of comment content, potentially through human or advanced LLM-based content labeling.

4.3 Exploratory Quantitative Evaluation of Bot Detection

Exploratory quantitative evaluation using simple machine learning classifiers provided a preliminary assessment of the heuristic-based bot detection approach.

- **Classification Model Performance:** Random Forest, SVM, and Neural Network classifiers were trained to distinguish between flagged and non-flagged accounts using the heuristic-derived features. The Random Forest classifier achieved the highest performance, with an accuracy of approximately 75% in cross-validation. SVM and Neural Network models showed slightly lower accuracy rates (around 70% and 68%, respectively).
- **Feature Importance:** Feature importance analysis from the Random Forest model indicated that posting frequency and content repetitiveness were the most influential features in distinguishing between flagged and non-flagged accounts. Account age, karma score, and em-dash presence also contributed to the model's predictive capability, albeit to a lesser extent. This reinforces the idea that behavioral patterns, captured by meta-metrics like posting frequency and content repetition, are more indicative of bot activity than simple content analysis of keywords or sentiment in this heuristic approach.
- **Confusion Matrices:** Figure 3 shows the confusion matrix for the Random Forest classifier, which achieved the best performance among the tested models. The model exhibited a tendency to misclassify some non-bot accounts as bots (false positives), while achieving relatively better performance in correctly identifying flagged accounts (true positives). This suggests that the heuristic-based approach, while effective in identifying some bot-like accounts based on meta-metrics, may also inadvertently flag some legitimate, highly active users.

Figure 3: Confusion Matrix for Random Forest Classifier

(Figure 3: Confusion Matrix for Random Forest Classifier is inserted here - See Appendix B, Figure 6 in the original notebook)

These quantitative results are preliminary and should be interpreted cautiously due to the lack of a true ground truth dataset and the limitations of relying solely on meta-metrics and basic content features. However, they provide initial evidence supporting the heuristic-based approach's potential for bot detection on Reddit and highlight areas for future refinement, particularly in reducing false positives and incorporating more sophisticated content-based features derived from labeled data.

4.4 Clustering Analysis

To further explore the underlying structure of the user-level data and visually assess potential groupings, dimensionality reduction and clustering techniques were applied. Principal Component Analysis (PCA) was used to reduce the dimensionality of the user-level dataset to two principal components, allowing for visualization in a 2D space. K-Means and DBSCAN

clustering algorithms were then applied to the PCA-transformed data to identify potential clusters of users based on their features.

Figure 4 displays scatter plots visualizing the results of K-Means and DBSCAN clustering, projected onto the first two principal components (PC1 and PC2).

Figure 4: K-Means and DBSCAN Clustering Visualizations

(Figure 4: K-Means and DBSCAN Clustering Visualizations are inserted here - See Appendix B, Figures 10 and 11 in the original notebook)

Note: The scatter plots visualize K-Means (top) and DBSCAN (bottom) clustering results after PCA dimensionality reduction. Points are colored by cluster assignment, and marker style indicates the 'Bot Category' assigned by heuristics. Axis labels indicate the top loading features for each Principal Component.

As shown in Figure 4, the visualizations reveal some degree of clustering in the user data, although distinct, well-separated clusters are not clearly evident.

- **K-Means Clustering:** The K-Means plot (top) shows a partitioning of the data into distinct clusters, but the clusters exhibit considerable overlap. PC1, representing a combination of features like average comment length, content cosine similarity, and comment length, is plotted on the x-axis. PC2, representing features like em-dash presence, karma, and median seconds between comments, is plotted on the y-axis. While K-Means forces data points into clusters, the visual overlap suggests that these clusters may not represent inherently distinct groups in terms of bot-like characteristics, and the algorithm may be imposing structure where clear separations do not naturally exist in the data based on these features alone.
- **DBSCAN Clustering:** The DBSCAN plot (bottom) identifies a central dense cluster (Cluster 0, shown in yellow) and designates a significant portion of data points as noise (Cluster -1, shown in purple), indicating that these points do not belong to any well-defined cluster based on DBSCAN's density-based criteria. Similar to K-Means, the axes represent PC1 and PC2 with the same feature loadings. The presence of a large noise cluster further supports the idea that clear, distinct groupings based on bot vs. human characteristics, as captured by these features and visualized in reduced dimensions, are not strongly present in the data. The limited cluster structure suggests that bot and human accounts, as identified by heuristics and visualized through PCA, may exist on a spectrum of behavioral characteristics rather than in clearly separated categories, at least based on the features used for clustering and visualization.

These clustering visualizations, while not revealing definitive bot clusters, visually reinforce the idea that distinguishing between bots and humans based solely on the meta-metrics and basic content features used in this study is a complex task. The lack of clear cluster separation suggests that more nuanced, potentially content-aware, features and more sophisticated analytical techniques might be necessary to effectively identify and categorize bot activity on Reddit beyond the initial heuristic flagging.

5 Discussion and Conclusion

This study explored the feasibility of using basic heuristic methods for detecting bot influence on Reddit discussions. The findings suggest that simple heuristics, based on readily observable account characteristics such as username patterns, account age, karma, posting frequency, content repetitiveness, and em-dash presence, can effectively identify a subset of accounts exhibiting bot-like behavior based on meta-metrics. Descriptive analysis of flagged accounts revealed patterns consistent with known bot characteristics, and exploratory quantitative evaluation using machine learning classifiers provided preliminary support for the predictive capability of the heuristic-based approach when relying on these meta-metrics and basic content features.

However, and importantly, the analysis of keyword frequencies and sentiment scores did *not* reveal statistically significant differences between flagged and non-flagged accounts. This key finding suggests that, while heuristic methods based on meta-metrics can identify certain bot-like accounts, these bots may be increasingly sophisticated in mimicking human language in terms of topical focus and sentiment expression, at least when analyzed using simple keyword frequency and sentiment analysis tools. This highlights a crucial limitation of relying solely on meta-metrics and basic content analysis for bot detection in the face of increasingly advanced automated accounts. It also suggests that simply analyzing keyword frequencies or overall sentiment polarity may not be sufficient to distinguish bots from humans in terms of content, and more nuanced, context-aware content analysis methods are needed.

The study also investigated potential narrative amplification and polarization associated with flagged accounts. While suggestive evidence of narrative amplification was observed through keyword analysis, and subtle differences in sentiment polarity were detected, the lack of statistical significance in these content-based analyses underscores the challenges in definitively attributing these phenomena to bot activity based on the methods employed here. LLM-generated word clouds, while thematically informative in visualizing subreddit topics, similarly did not provide clear differentiation between bot and human language use in terms of content.

In essence, the study demonstrates that while basic heuristics can flag accounts exhibiting bot-like behavior, these heuristics alone are insufficient to definitively identify bots or understand their influence on content without more advanced content-based analysis methods and labeled data.

5.1 Limitations

This study has several limitations that should be considered when interpreting the findings:

- **Heuristic Accuracy and Reliance on Meta-metrics:** The heuristic-based bot detection method, while interpretable and readily implementable, is inherently limited in its accuracy, particularly when relying solely on meta-metrics. It relies on predefined rules that may not capture the full spectrum of bot behaviors, especially increasingly sophisticated bots that closely mimic human users and whose content may be indistinguishable from human content when analyzed with basic methods[1]. The

exploratory quantitative evaluation revealed a tendency for false positives, indicating that the heuristics may inadvertently flag some legitimate users as bots. Crucially, the lack of significant differences in keyword frequencies and sentiment scores suggests that meta-metrics alone are insufficient for a comprehensive and nuanced understanding of bot activity, and that content-based analysis is essential for future progress.

- **Lack of Ground Truth and Content Labeling:** The absence of a true "ground truth" dataset for bot identification on Reddit poses a significant challenge for rigorous quantitative evaluation. Moreover, the study did not incorporate a crucial element for content-based bot detection: labeled data where comment bodies are categorized as bot-generated or human-generated. The exploratory machine learning evaluation provides only a preliminary indication of the heuristic approach's predictive capability based on meta-metrics but cannot definitively assess its accuracy against a verified bot/human classification that incorporates content analysis. Future research must prioritize the creation of such labeled datasets, potentially through human annotation or LLM-assisted labeling, to enable the development of more robust content-aware bot detection models.
- **Data Sampling:** The data collection was limited to the top 50 hottest posts per subreddit and a maximum of 2000 comments per post. This sampling approach may not capture the full diversity of discussions and bot activity across Reddit. Furthermore, the analysis focused on only the top 5 subreddits, limiting the generalizability of the findings to the entire Reddit platform.
- **LLM Dependency for Exploratory Analysis:** The study's reliance on Google Gemini for keyword extraction and sentiment summarization, while providing exploratory insights, introduces a dependency on a specific LLM API and is not a substitute for direct content-based bot detection. While efforts were made to ensure reproducibility by tuning LLM parameters, the results should be seen as qualitative explorations rather than definitive quantitative measures of bot influence on content.

5.2 Future Research Directions

Future research should address these limitations and further explore the detection of bots on Reddit, with a particular emphasis on moving beyond meta-metrics and incorporating content-based analysis:

- **Refinement of Heuristics and Integration of Content Features:** Further research is needed to refine and expand the heuristic-based bot detection method, integrating content-based features derived from labeled data. This could involve incorporating more sophisticated meta-metrics, such as network-based metrics derived from user interaction patterns within subreddits, but crucially, it should also prioritize features extracted directly from the comment text itself, leveraging human-validated labeled data to identify nuanced linguistic patterns indicative of bot-generated content, going beyond simple metrics like em-dash presence to encompass stylistic features, semantic coherence, contextual relevance, and even subtle cues in argumentation and conversational style. Adaptive heuristics that dynamically adjust to evolving bot behaviors, including content generation strategies, could also be explored.

- **Development of Ground Truth Datasets with Human-Validated Content Labels:** A critical step for future research is the development of high-quality, labeled datasets for Reddit bot detection that include human-validated labels for comment content. This could involve combining manual labeling by expert Reddit moderators with LLM-assisted labeling techniques to categorize comment bodies as bot or human-generated, with a rigorous human review and validation process to ensure label accuracy and minimize biases. This labeled data is essential to train and evaluate models that can effectively leverage content-based features for bot detection, moving beyond the limitations of meta-metrics and basic content features.
- **Advanced Machine Learning Models for Content-Aware Bot Detection:** Future studies should investigate the application of more advanced machine learning models, including deep learning architectures and graph neural networks, trained on datasets with human-validated content labels, for Reddit bot detection. These models, capable of processing and understanding natural language, may be better suited to capture the complex behavioral and linguistic patterns of sophisticated bots and potentially improve detection accuracy by learning directly from the content of bot and human comments, going beyond simple keyword or sentiment analysis to understand deeper semantic, stylistic, and contextual cues that differentiate human and bot-generated text.
- **Hybrid Approaches Combining Meta-metrics and Content Analysis:** Combining heuristic-based methods, focused on meta-metrics, with machine learning models trained on content-labeled data could offer a promising avenue for future research. Heuristics could be used for initial feature engineering and data pre-processing of metadata, while machine learning models could be trained to learn more complex patterns and improve detection accuracy by incorporating both meta-metric and content-based features derived from human-validated labeled data, creating a more robust and nuanced bot detection system that leverages the strengths of both rule-based and data-driven approaches.
- **Real-time Bot Detection Systems with Content Analysis Capabilities and Ethical Considerations:** Developing real-time bot detection systems that can proactively identify and flag malicious activity on Reddit is a crucial direction for future work. This would require efficient and scalable algorithms capable of processing large volumes of streaming Reddit data and incorporating sophisticated content analysis in real-time, potentially through optimized and ethically vetted LLM-based feature extraction or lightweight content classification models, while also carefully addressing the computational cost, latency considerations, and ethical implications of real-time content processing and automated bot flagging. Ethical considerations, particularly regarding potential biases in content labels and algorithmic detection, and the need for transparency and user recourse mechanisms in content-based bot detection systems, must be central to future development.

5.3 Conclusion

This study provides a preliminary exploration into the detection of bots on Reddit using basic heuristic methods and LLMs for exploratory analysis. While heuristic-based approaches offer a readily implementable and interpretable starting point for identifying some bot-like accounts based on meta-metrics, the study highlights the crucial limitation of relying solely on meta-metrics and basic content analysis in the face of increasingly sophisticated bots. The key takeaway is the necessity of moving beyond meta-metrics and incorporating human-validated content-based analysis, which requires the development of labeled datasets and the application of advanced machine learning models capable of understanding and learning from the textual content of Reddit comments, while carefully considering ethical implications and ensuring fairness and transparency. Future research must prioritize these directions to develop more robust, accurate, ethically sound, and content-aware bot detection systems for this complex and evolving online environment. The ongoing effort to detect and mitigate bot influence, particularly through advanced content-aware methods, is crucial for maintaining the integrity and trustworthiness of online social platforms and ensuring a healthy digital public sphere.

References

1. Hurtado S., Ray P., & Marculescu R. (2019). Bot detection in reddit political discussion. *ResearchGate*. https://www.researchgate.net/publication/332340547_Bot_Detection_in_Reddit_Political_Discussion
2. Dawson A. (2024). Does reddit have a bot problem? absolutely. In *Lunio*. https://www.lunio.ai/blog/reddit-bots
3. R/botwatch. (2022). In *Reddit*. https://www.reddit.com/r/botwatch/
4. Ng L. H. X., & Carley K. M. (2022). Assembling a multi-platform ensemble social bot detector with applications to US 2020 elections. *arXiv*. https://arxiv.org/html/2401.14607v1
5. Kutlu M., & Selçuk A. A. (2025). Evaluation of social bot detection models. *ResearchGate*. https://www.researchgate.net/publication/361038547_Evaluation_of_social_bot_detection_models
6. Tang J. (2024). *Lessons learned from scaling up cloudflare's anomaly detection platform*. https://blog.cloudflare.com/lessons-learned-from-scaling-up-cloudflare-anomaly-detection-platform/
7. Tang J. (2024). How does anomaly detection work? : R/f5networks. In *Reddit*. https://www.reddit.com/r/f5networks/comments/13jw4qg/how_does_anomaly_detection_work/
8. Damasceno R. (2019). Identify-bots-reddit-comment-network: Characterization and classification of bots using only structural characteristics of the network. In *GitHub*. https://github.com/DamascenoRafael/identify-bots-reddit-comment-network
9. Skowronski J. (2019). Identifying trolls and bots on reddit with machine learning. In *Medium*. https://medium.com/towards-data-science/identifying-trolls-and-bots-on-reddit-with-machine-learning-709da5970af1
10. Baumgartner J., Zannettou S., Keegan B., Squire M., & Blackburn J. (2020). The pushshift reddit dataset. *AAAI Publications*. https://ojs.aaai.org/index.php/ICWSM/article/view/7347/7201
11. Tkachenko V. (2021). Reddit comments dataset. In *ClickHouse Docs*. https://clickhouse.com/docs/getting-started/example-datasets/reddit-comments
12. Tourond M. (2019). Reddit-bot-detector: A python bot that detects reddit bots. In *GitHub*. https://github.com/MatthewTourond/Reddit-Bot-Detector
13. Gillham J., & Lambert M. (2023). *Are you getting advice from a human or bot? Reddit shows spikes in AI content*. https://originality.ai/blog/reddit-shows-spikes-in-ai-content
14. Bush C. (2023). *The em dash and AI: A conjunction - night water*. https://www.nightwater.email/em-dash-ai/
15. How to identify bots on reddit : R/LearnUselessTalents. (2023). In *Reddit*. https://www.reddit.com/r/LearnUselessTalents/comments/15tzjkb/how_to_identify_bots_on_reddit/
16. u/tyrannosnorlax. (2022). Bots. How to identify them, and why do they exist on reddit? In *Reddit*. https://www.reddit.com/user/tyrannosnorlax/comments/t0h466/bots_how_to_identify_them_and_why_do_they_exist/
17. Bot problem - how to identify bot accounts (99% accuracy) : R/7daystodie. (2025). In *Reddit*. https://www.reddit.com/r/7daystodie/comments/1au1g1s/bot_problem_how_to_identify_bot_accounts_99/

Appendices

Keyword Generation and Subreddit Selection Details

```
# Fetch a large subset of popular subreddits (large limit makes this
 ↪  representative of the largest overall subreddits by subscribers, check:
 ↪  https://gummysearch.com/tools/top-subreddits/)
subreddits = list(reddit.subreddits.popular(limit=1000))

# Create a DataFrame using list comprehension for better performance
subs_df = pd.DataFrame([{
    "Name": subreddit.display_name,
    "Subscribers": subreddit.subscribers,
    "Description": subreddit.public_description,
    "Over 18": subreddit.over18,
    "Submission Type": subreddit.submission_type
} for subreddit in subreddits]).sort_values(by="Subscribers",
 ↪  ascending=False, ignore_index=True)

# Print the top 10
subs_df.head(10)
```

	Name	Subscribers	...	Over 18	Submission Type
0	funny	66604854	...	False	any
1	AskReddit	53169243	...	False	self
2	gaming	46005690	...	False	any
3	worldnews	44847466	...	False	link
4	todayilearned	40125637	...	False	link
5	aww	37645187	...	False	link
6	Music	36991965	...	False	any
7	memes	35397966	...	False	link
8	movies	34851952	...	False	any
9	Showerthoughts	34152969	...	False	self

```
import ast

response = generate("What are some keywords I can use to create a list of
 ↪  subreddits which are likely to be influenced by bots because of their
 ↪  controversial nature? These are keywords that I would look for within a
 ↪  subreddit's name or description. For example: \"news\", \"politics\",
 ↪  \"discussion\", \"war\", \"vaccines\", \"controversial\", \"conflict\",
 ↪  etc.\n\nKeep the answer short, only including 50 keywords and saving
 ↪  them in a python list as follows [\"key1\",\"key2\",...]. Send the
 ↪  output as text not as code.")
```

```
bot_influence_keywords = ast.literal_eval(response.replace("\n", ""))

for i in range(0, len(bot_influence_keywords), 5):
    print(*bot_influence_keywords[i:i+5])
```

```
news politics discussion war vaccines
controversial conflict debate election government
russia china ukraine israel palestine
climate immigration guns religion socialism
capitalism feminism lgbt transgender race
identity activism protest censorship freedom
rights justice police crime law
economy finance markets technology science
health global world opinion truth
facts bias propaganda conspiracy agenda
```

```
# Score subreddits based on subscribers and keywords in description
def calculate_bot_influence_score(row):
    score = 0

    # Large subscriber base increases potential for bot activity
    if row['Subscribers'] > 10000000:
        score += 5
    elif row['Subscribers'] > 5000000:
        score += 4
    elif row['Subscribers'] > 1000000:
        score += 3

    # Check for keywords in description and subreddit name
    description = row['Description'].lower()
    sub_name = row['Name'].lower()
    for keyword in bot_influence_keywords:
        if keyword in description:
            score += 1
        if keyword in sub_name:
            score += 1

    return score

subs_df['Bot Score'] = subs_df.apply(calculate_bot_influence_score, axis=1)

# Get top 50 most vulnerable subreddits
top_vulnerable = subs_df.nlargest(50, 'Bot Score')[['Name', 'Subscribers',
 ↪  'Submission Type', 'Bot Score']].reset_index(drop=True)
top_vulnerable.head(10)
```

	Name	Subscribers	Submission Type	Bot Score
0	worldnews	44847466	link	9
1	technology	18520150	link	9
2	IndiaSpeaks	1049211	any	9
3	news	29802412	link	8
4	pcmasterrace	14764738	any	8
5	politics	8785328	link	8
6	movies	34851952	any	7
7	science	33804429	link	7
8	askscience	26054520	self	7
9	AmItheAsshole	24117454	self	7

Figures and Tables

Figure 1: Word Cloud for /r/politics *(See Appendix B, Figure 1 in the original notebook)*

Figure 2: Word Cloud for /r/worldnews *(See Appendix B, Figure 2 in the original notebook)*

Figure 3: Word Cloud for /r/news *(See Appendix B, Figure 3 in the original notebook)*

Figure 4: Word Cloud for /r/technology *(See Appendix B, Figure 4 in the original notebook)*

Figure 5: Word Cloud for /r/science *(See Appendix B, Figure 5 in the original notebook)*

Figure 6: Confusion Matrix for SVM Classifier *(See Appendix B, Figure 7 in the original notebook)*

Figure 7: Confusion Matrix for Neural Network Classifier *(See Appendix B, Figure 8 in the original notebook)*