Is a human being on the other side?

1 Problem Description

As technology develops, bots are becoming more and more sophisticated and difficult to distinguish from humans. Their ability to mimic human overtones and characteristics of posts and comments poses a serious challenge. Artificially generated content can displace human activity on social media.

Research question: Are we able to tell if a post or comment (publication) was written by a bot?

Purpose: To analyze the similarity of Reddit publications created by bots as well as humans.

2 How did we get data?

User contributions from the five most popular subreddits on the Reddit platform were used as a data source. The data was collected using a custom script that allowed direct access to the data using the PRAW library and the platform's programming interface (API).

To conduct the analysis, a comprehensive dataset was collected from the five most popular subreddits on the Reddit platform: **r/funny, r/AskReddit, r/gaming, r/worldnews** and **r/todayilearned**. The data collection process, including both posts, comments and user information, lasted continuously for 6 days, resulting in the creation of a comprehensive dataset. Including Reddit-tagged **official bots** and "human" users along with **unclassified bots**.

This is how we managed to collect the following number of data:

comments: 800 000

posts: 4 500

users (including unclassified bots): 370 000

classified bots (tagged by Reddit): 10 000

3 How to describe the bots?

Based on a literature review and an analysis of the characteristics of Reddit platform users, we created 16 key indicators to identify bots. Indicators based on text analysis were based on comparison of text representation obtained with *TfidfVectorizer* model. The indicators were divided into 4 groups (For more details, please visit our repository):

- **Activity:** Bots are distinguished by atypical activity rating values (*link_karma*, *comment_karma*), young account age (*account_age*), and a specific ratio of comments to posts (*comment_post_ratio*).
- **Language style and interactions:** Generate repetitive content (*avg_cosine_similarity*, *all_users_similarity*, *parent_child_similarity*), which were calculated using a cosine similarity measure, are characterized by extreme comment lengths (*avg_comment_length*, *min_comment_length*, *max_comment_length*), and a certain depth of threads in which they participate (*avg_thread_depth*).
- **Linguistic complexity:** They create texts with low diversity (*avg_ttr*), which was calculated using the type-toke-ratio metric, specific linguistic complexity and text readability (*avg_flesch_kincaid_grade*), calculated using the flesch-kincaid grade metric, and high pattern repeatability (*ngram_overlap*), calculated using the ngram overlap technique.
- **Engagement:** They get unusual scores on the platform (*avg_score*) and fewer responses to comments (*avg_num_replies*).

Learn more at the project repository (GitHub)

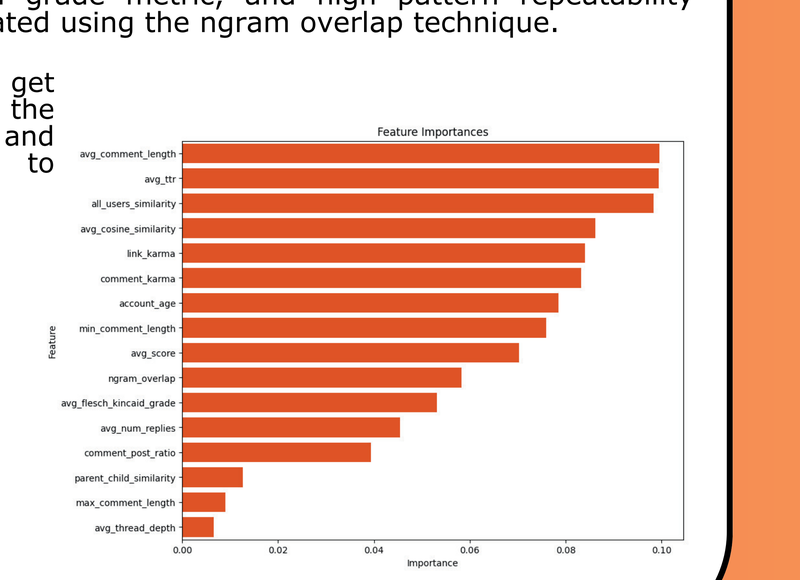

Figure 1: Significance of features as determined by the RandomForest model.

4 Are they similar to us?

Analysis using *Random Forest*, *CatBoost* and *SVM* classifiers showed significant difficulty in assigning labels achieving the best results at the *F1* metric level of 0.58. These results may have been caused by the blurring of the decision boundary and the presence of unclassified bots in the user class. For this reason, the analysis was deepened by using single-class classifiers such as *OneClassSVM* and *LocalOutlierFactor*, which also achieved a classification quality of *F1* equal to 0.6.

The next step of the in-depth analysis was to visualize the data using the nonlinear dimensionality reduction method ***UMAP***. The projection of a multidimensional space onto a two-dimensional space, shows the low separability of classes (official bots and users along with unclassified bots) for the distribution of features becomes noticeable, suggesting their similarity.

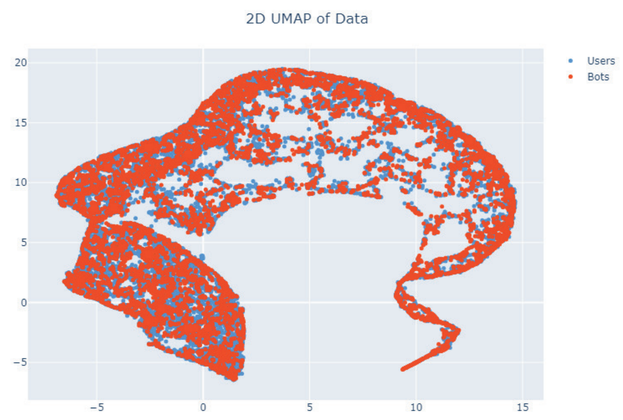

Figure 2: UMAP 2D for User and Bot data.

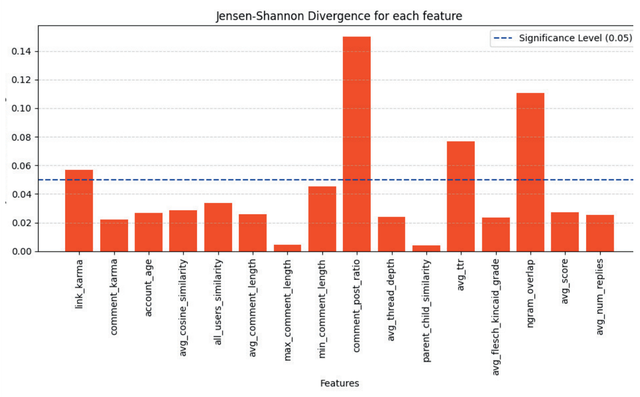

Figure 3: Results of the Jensen-Shannon divergence analysis for all considered features.

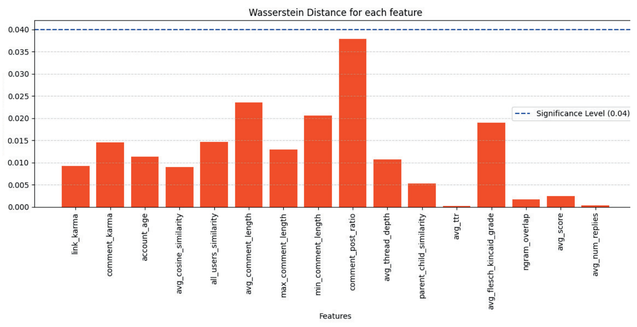

Figure 4: Results of Wasserstein distance analysis for all considered features.

To measure the difference in feature distributions between bots and users, we used **Jensen-Shannon divergence**, which measures the similarities between two probability distributions.

In the results, higher values indicate greater differences between the distributions for a given characteristic. On the graph, key differences were observed for *min_comment_length* and *comment_post_ratio*, suggesting that post length and number of comments best differentiate bots from humans. Overall, the results for most traits are very small (less than 0.05) suggesting that the probability distributions are very similar to each other.

Then, to prove the similarity of the distributions, the **Wasserstein distance** measure was used, which measures the minimum cost of transforming one distribution into another. The distances obtained for all traits were below the 0.04 threshold, suggesting that their distributions are very similar in both groups.

5 Anxious future of social media

In our research on detecting untagged bots among ordinary users using analysis of various parameters as well as approaches, we concluded that **bots are almost indistinguishable from humans**. This situation poses a serious problem for the future of digital communication and online security.

The lack of clear patterns that distinguish bots from users informs us that **bots have learned to write in a "human way" extremely effectively.** In our opinion, this similarity is conditioned by the fact that with the development of generative technology, bots are increasingly using large-scale language models (LLMs) that were learned on human publications. As a result, the difference in the way bots and humans communicate is becoming virtually imperceptible. This can lead to:

Information manipulation: Bots can be used to spread disinformation en masse in ways that are difficult to detect.

Decreases in social trust: People's lack of trust in online interactions, suspecting that the interlocutor may be a bot, rather than a human.

Challenges in content moderation: Automated algorithms and manual moderation can be insufficient and ineffective in identifying bots.

Katedra Sztucznej Inteligencji

Politechnika Wrocławska

The project made in the class "Digital Media Analysis" during the course Artificial Intelligence 2024.

Authors: *Dawid Kopeć, Dawid Krutul, Maciej Wizerkaniuk*