A Proposal for Uncovering Hidden Social Bots via Genetic Similarity

Edoardo Allegrini[1], Edoardo Di Paolo[1], Marinella Petrocchi[2,3] and Angelo Spognardi[1]

[1]Computer Science Department, Sapienza University of Rome, Italy
[2]Istituto di Informatica e Telematica, CNR, Pisa, Italy
[3]Scuola IMT Alti Studi Lucca, Italy

Abstract

Social media platforms face an ongoing challenge in combating the proliferation of social bots, automated accounts that are also known to distort public opinion and support the spread of disinformation. Over the years, social bots have evolved greatly, often becoming indistinguishable from real users, and more recently, families of bots have been identified that are powered by Large Language Models to produce content for posting. We suggest an idea to classify social users as bots or not using genetic similarity algorithms. These algorithms provide an adaptive method for analyzing user behavior, allowing for the continuous evolution of detection criteria in response to the ever-changing tactics of social bots. Our proposal involves an initial clustering of social users into distinct macro species based on the similarities of their timelines. Macro species are then classified as either bot or genuine based on genetic characteristics. The preliminary idea we present, once fully developed, will allow existing detection applications based on timeline equality alone to be extended to detect bots. By incorporating new metrics, our approach will systematically classify non-trivial accounts into appropriate categories, effectively peeling back layers to reveal non-obvious species.

Keywords

Social bot detection, Bioinformatics, Social Network

1. Introduction

The digital age has brought with it an unprecedented proliferation of accounts on social platforms, resulting in a diverse and complex ecosystem. Of particular note are automated accounts, commonly known as social bots [1, 2]. These digital artifacts have received increasing attention, not only because of their ubiquity, but also because of the role they often play as vehicles for misinformation and propaganda [3]. Recently, families of bots have been heuristically identified that use large language models to produce content for publication. The proliferation of these advanced bots raises concerns about the inability of researchers to detect them [4].

One of the most important strands of research began with the realization that bots, programmed to pursue specific goals, often operate in a coordinated manner and exhibit similar behaviors. In particular, one modeling and detection technology that was notably relevant was that based on digital DNA [5, 6]. Digital DNA is a string of characters, each of which associated with a specific account action, representing the timeline of the account. This modeling technique has been used in several studies, see, e.g., [7, 8, 9, 10].

Based on the intuition that accounts of the same type behave similarly, if not exactly the same, we propose a method for classifying bot accounts that may be mistaken for real accounts.

Discovery Science - Late Breaking Contributions 2024
allegrini@di.uniroma1.it (E. Allegrini); dipaolo@di.uniroma1.it (E. Di Paolo); marinella.petrocchi@iit.cnr.it (M. Petrocchi); spognardi@di.uniroma1.it (A. Spognardi)
0009-0003-8842-6873 (E. Allegrini); 0000-0001-9216-8430 (E. Di Paolo); 0000-0003-0591-877X (M. Petrocchi); 0000-0001-6935-0701 (A. Spognardi)

2. Proposed Method

The classification approach we propose has been designed for Twitter/X users, but minor adjustments can easily adapt it for use on other social media platforms. Figure 1 shows the scheme of the procedure.

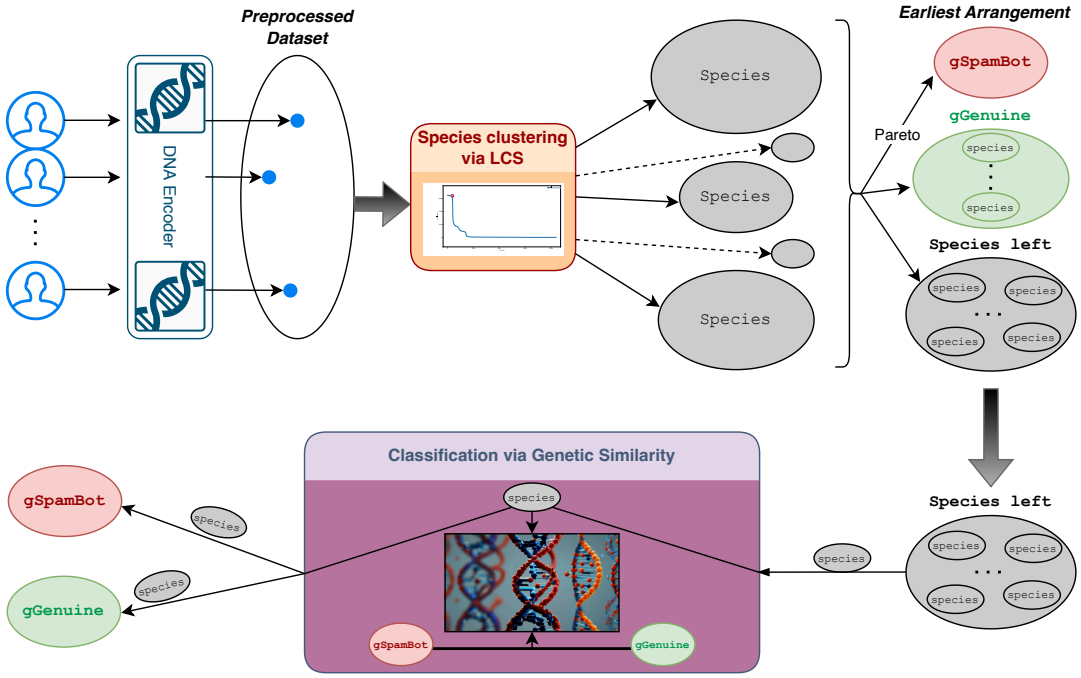

Figure 1: Steps for hidden social bots classification

Digital DNA The first phase involves pre-processing users by encoding their online behavior using Digital DNA. Digital DNA aims to compactly represent the behavior of a social account, using a sequence of characters from an alphabet $\mathbb{B}$, such as the following made of three characters:

$$\mathbb{B}^3_{type} = \begin{Bmatrix} A \rightarrow \text{plain tweet} \\ T \rightarrow \text{retweet} \\ C \rightarrow \text{reply} \end{Bmatrix} \tag{1}$$

Clustering users into species After the pre-processing stage, users are grouped into macro species, based on the concept of the Longest Common Substring (LCS) [11]. The LCS of two or more strings is the longest string that is a substring of all of them. Figure 2 shows an example of an LCS curve, where the abscissae are the groups of k users, and the ordinate is the length of the LCS for each of the user groups. Considering a dataset of N users paired with their digital DNA sequences, the LCS computation is performed in linear time [11] between k users, where $k \in \{2, ..., N\}$. LCS is an indicator of behavioral similarity within the user group: When the sub-DNAs in the LCS curve are of approximately constant length, we can deduce that the users with these sequences have similar behavior. Conversely, if we observe a significant drop in the LCS curve, we know that the behavior of newly added users differs considerably from that of the users in the previous group. As an example, the red circle in the figure marks the point where the curve exhibits a significant drop. To identify and cluster users into species, the first significant drop in the LCS curve is detected, which marks a behavioral shift. Users associated with that drop are grouped into a new species and subsequently removed. The LCS curve is then recomputed for the remaining users that have not been assigned to a cluster yet. This process is repeated, progressively segmenting and grouping all users in the dataset into distinct clusters called *species*.

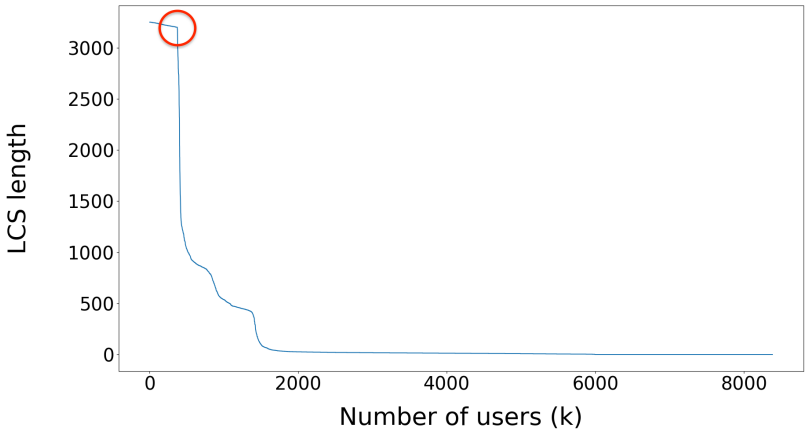

Figure 2: Plot of an illustrative LCS curve

Spambot and genuine accounts: first arrangement From the macro clusters obtained, two key groups are constructed: *gSpamBot*, by selecting among the species the one demonstrating evident bot-like behavior, and *gGenuine*, with a high predominance of genuine users. Later, these two groups will be populated by the remaining users, by adopting specific algorithms developed to measure the genetic similarity.

To establish the initial *gSpamBot* and *gGenuine* groups, the following idea is applied: the LCS of a species represents the users within it; therefore, a group with a long LCS indicates similar social behavior, possibly an indicator of social bots. On the other side, a group with a short LCS implies a diversity in the social behavior of its members. The construction of the initial *gSpamBot* group is inspired by the Pareto principle, which seeks to determine the subset of individuals that have the most significant impact on the overall community. Based on the original dataset, users have now been categorized into three groups: *gSpamBot*, which comprises a considerable number of users exhibiting very similar behavior, strongly indicating that they are social bots; *gGenuine*, formed by merging species that demonstrate human-like actions; and lastly, species that do not fall into either *gSpamBot* or *gGenuine* and are therefore unlabeled at this stage.

Classification of species using genetic similarity To classify the unlabeled species (those colored gray in Figure 1), whose users are not immediately classifiable as bots or not, we propose an algorithm that uses custom genetic similarity metrics. The idea is using a sequence alignment algorithm —a well-established technique in bioinformatics— to compare the LCS of each unlabeled species with those of the two primary groups. After this alignment, our algorithm introduces a structured classification process. The process involves two key steps: first, calculating a similarity score based on the alignment of LCS sequences between species; and second, evaluating a new metric that considers the relative similarity of DNA sequences within a species and the size of the population contributing to that similarity. By integrating these procedures, we believe that our approach can effectively label the previously unlabeled species as either *gSpamBot* or *gGenuine*, thereby completing the account classification process.

3. Conclusions

In this short paper, we presented the idea of a new approach to social bot detection that we hope will not only achieve effective classification, but also maintain a transparent decision-making process.

We plan to implement and test our proposed classifier using well-established bot repositories (see the ones published on the site of the OSOME research unit at Indiana University) as well as recently discovered datasets where social bots use Large Language Models (LLMs) to write their posts.

Acknowledgments

This work is partially supported by project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU; by project re-DESIRE (DissEmination of ScIentific REsults 2.0), funded by IIT-CNR; by project 'Prebunking: predicting and mitigating coordinated inauthentic behaviors in social media', funded by Sapienza University of Rome.

References

[1] E. Ferrara, O. Varol, C. Davis, F. Menczer, A. Flammini, The rise of social bots, Communications of the ACM 59 (2016) 96–104.
[2] S. Cresci, A decade of social bot detection, Communications of the ACM 63 (2020) 72–83.
[3] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, F. Menczer, The spread of low-credibility content by social bots, Nature Communications 9 (2018) 1–9.
[4] K.-C. Yang, F. Menczer, Anatomy of an ai-powered malicious social botnet, Journal of Quantitative Description: Digital Media 4 (2024).
[5] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi, Dna-inspired online behavioral modeling and its application to spambot detection, IEEE Intell. Syst. 31 (2016) 58–64.
[6] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi, Social fingerprinting: Detection of spambot groups through dna-inspired behavioral modeling, IEEE Transactions on Dependable and Secure Computing 15 (2018) 561–576.
[7] E. Di Paolo, M. Petrocchi, A. Spognardi, From online behaviours to images: A novel approach to social bot detection, in: Computational Science, 2023, pp. 593–607.
[8] N. Pasricha, C. Hayes, Detecting bot behaviour in social media using digital dna compression, in: Artificial Intelligence and Cognitive Science, 2019.
[9] R. Gilmary, A. Venkatesan, Entropy-based automation detection on twitter using dna profiling, SN Computer Science 4 (2023) 847.
[10] V. Chawla, Y. Kapoor, A hybrid framework for bot detection on twitter: Fusing digital dna with bert, Multimedia Tools and Applications 82 (2023) 30831–30854.
[11] M. Arnold, E. Ohlebusch, Linear time algorithms for generalizations of the longest common substring problem, Algorithmica 60 (2011) 806–818.