BotSentinel

Advisor: João Almeida (joao.rafael.almeida@ua.pt)

1. Background

In recent years, the rise of Artificial Intelligences capable of generating coherent text, realistic images, and, more recently, interacting on social media has sparked growing debate about the authenticity of online profiles. Twitter, one of the most influential social platforms, has become a fertile ground for AI-driven bots, which can be used to spread misinformation, manipulate discussions, and create artificial engagement. The ease with which an automated profile can mimic human behavior raises concerns about the reliability of information and the formation of genuine opinions in digital spaces.

The Dead Internet Theory serves as an inspiration for this project. It suggests that much of the online content we consume may be artificially generated rather than produced by real people. According to this theory, the internet as we know it is largely dominated by bots and automated algorithms, creating an illusion of human participation and engagement. Over time, humans have increasingly interacted with automated profiles without realizing that these interactions are not genuine. This shift could lead to a loss of authenticity, where real human voices are replaced by algorithms that shape behaviors and opinions. While this idea might sound exaggerated or like a conspiracy theory, the reality is that we are moving in that direction. Today's internet is far more controlled and homogeneous than the chaotic and unpredictable web of the past. The once-diverse ecosystem of user-created small websites has been replaced by a handful of large platforms dominated by corporations, prioritizing the monetization of interactions—often at the expense of user experience. We are not yet fully immersed in the dystopia described by the theory, but the signs of this transformation are becoming increasingly evident, making it crucial to question and understand its implications.

This project aims to develop a system that enables users to identify profiles suspected of being AI-driven, leveraging a combination of community-based evaluations. The goal is to provide users with tools to actively contribute to detecting and classifying bots, helping to increase transparency in online interactions and ensuring that digital participation is not covertly manipulated by AI.

2. Project Objectives

The project aims to develop an innovative solution for identifying accounts managed by artificial intelligence (AI). While existing tools address this issue in a general way, this project seeks to provide direct and precise insights into specific accounts. To achieve this, the key objectives are:

- Develop an accessible and flexible platform, available as both a website and a browser plug-in. The plug-in will automatically display an account’s credibility while browsing, while the website will allow users to analyze profiles through a simple search, making it ideal for mobile devices.

- Implement a Twitter authentication system to ensure that only verified users can vote. This feature will enhance the credibility of evaluations and enable a more accurate analysis of whether an account is AI-operated.
- Create an intuitive and informative service, allowing users to obtain clear and immediate results about profile credibility without requiring technical knowledge or complex procedures.
- Establish a collaborative evaluation system, where votes from authenticated users will be combined with proprietary algorithms to determine the likelihood of an account being AI-managed. This approach will integrate collective intelligence with technical analysis.
- Introduce a credibility mechanism linked to phone numbers. Users will have the option to link their number, which will be publicly visible to enhance transparency and trust in their profile without affecting their credibility score. Each number will be limited to a single account.
- Assign different weights to votes based on user credibility. Votes from accounts with a lower probability of being AI will have greater influence, while those from suspicious accounts will have reduced impact, ensuring fairness and integrity in evaluations.
- Ensure public access to general information, regardless of authentication. However, voting will be restricted to verified users to maintain process reliability.
- Reduce misinformation and unwanted content on social media by providing users with tools to identify suspicious accounts and make more informed decisions.
- Enable users to view evaluations and comments on specific profiles, promoting transparency and providing additional context on an account's reputation.
- Implement a binary voting system ('Yes' or 'No'), with an option to add comments explaining the evaluation. This will simplify the process while maintaining analytical depth.

3. Workplan

The work plan can be divided into the following tasks:

1. Investigate the use of AI in creating and managing social media accounts.
   - Requirements Analysis
   - Define system requirements, user personas, and usage scenarios.
2. System architecture and design
   - Select the technologies to be used and define how they will interact within the system.
3. Development and implementation
   - Develop the web service.
   - Build a browser extension.
4. System validation
   - Conduct unit and integration tests to ensure the correct functioning of system components.