UNIVERSITY OF LJUBLJANA
SCHOOL OF ECONOMICS AND BUSINESS

MASTER'S THESIS

THE DEAD INTERNET THEORY: DETECTING FAKE PROFILES ON SOCIAL MEDIA USING MACHINE LEARNING

Ljubljana, January 2026 JOHANNES FRÄNKLE

AUTHORSHIP STATEMENT

The undersigned Johannes Fränkle, a student at the University of Ljubljana, School of Economics and Business, (hereafter: UL SEB), author of this written final work of studies with the title 'The Dead Internet Theory: Detecting fake profiles on social media using machine learning', prepared in collaboration with mentor Luka Tomat, Ph. D and co-mentor Prof. Dr. Frank Morelli.

D E C L A R E

1. this written final work of studies to be based on the results of my own research;

2. the printed form of this written final work of studies to be identical to its electronic form;

3. the text of this written final work of studies to be language-edited and technically in adherence with the UL SEB Technical Guidelines for Written Works, which means that I cited and / or quoted works and opinions of other authors in this written final work of studies in accordance with the UL SEB Technical Guidelines for Written Works;

4. to be aware of the fact that plagiarism (in written or graphical form) is a criminal offence and can be prosecuted in accordance with the Criminal Code of the Republic of Slovenia;

5. to be aware of the consequences a proven plagiarism charge based on the this written final work could have for my status at the UL SEB in accordance with the relevant UL SEB Rules;

6. to have obtained all the necessary permits to use the data and works of other authors which are (in written or graphical form) referred to in this written final work of studies and to have clearly marked them;

7. to have acted in accordance with ethical principles during the preparation of this written final work of studies and to have, where necessary, obtained permission of the Ethics Committee;

8. my consent to use the electronic form of this written final work of studies for the detection of content similarity with other written works, using similarity detection software that is connected with the UL SEB Study Information System;

9. to transfer to the University of Ljubljana free of charge, non-exclusively, geographically and time-wise unlimited the right of saving this written final work of studies in the electronic form, the right of its reproduction, as well as the right of making this written final work of studies available to the public on the World Wide Web via the Repository of the University of Ljubljana;

10. my consent to publication of my personal data that are included in this written final work of studies and in this declaration, when this written final work of studies is published;

11. that I have verified the authenticity of the information derived from the records using artificial intelligence tools.

Ljubljana, November 3th, 2025 Author's signature: 

ABSTRACT

Social media platforms play a central role in shaping digital identity and online communication, but they are increasingly affected by automated accounts and synthetic activity. Building on the Dead Internet Theory (DIT), which argues that much online interaction is artificially generated, this thesis investigates bot detection on Instagram, a platform that has received comparatively little academic attention despite its global relevance.

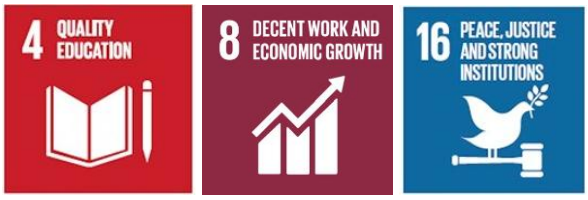
The research applies the Design Science Research (DSR) framework in combination with the CRISP-DM process model to ensure both methodological rigor and practical relevance. Data was collected through automated scraping and incorporates three modalities: profile metrics, textual comments, and visual features. Following Georges' identity layer model, these were mapped to declarative, active, and calculated aspects of digital identity. Due to the scarcity of labeled datasets, a semi-supervised learning approach was adopted, leveraging a small labeled set alongside a larger pool of unlabeled accounts.

Several machine learning models were trained and evaluated, including Random Forest, Support Vector Machines, Logistic Regression, and LightGBM. Results show strong performance for models trained on profile metrics, with Random Forest achieving 97% accuracy. In contrast, models using visual features performed significantly worse. When applied to the larger unlabeled dataset, Random Forest labeled around 2% of accounts as bots, while LightGBM identified nearly 18%, reflecting different sensitivities and assumptions in the models. The semi-supervised approach expanded the training dataset from 634 to 9,562 labeled instances, labeling 89.3% of the unlabeled data with high confidence. This demonstrates the method's effectiveness in low-labeled-data environments common to Instagram research.

This thesis contributes by demonstrating the potential of multimodal, semi-supervised approaches for bot detection on Instagram. It also highlights methodological challenges, such as limited labeled data and model interpretability, while outlining directions for future research, including neural networks, sentiment analysis, and reverse image verification.

KEY WORDS: social media, machine learning semi-supervised learning, Design Science Research, CRISP-DM

SUSTAINABLE DEVELOPMENT GOALS



POVZETEK

Platforme družbenih medijev igrajo osrednjo vlogo pri oblikovanju digitalne identitete in spletne komunikacije, vendar nanje vse bolj vplivajo avtomatizirani računi in sintetične dejavnosti. Na podlagi teorije mrtvega interneta (angl. Dead Internet Theory - DIT), ki trdi, da je veliko spletnih interakcij ustvarjenih umetno, to magistrsko delo preučuje odkrivanje robotov na Instagramu, platformi, ki je kljub svoji globalni pomembnosti prejela sorazmerno malo akademskega zanimanja. Raziskava uporablja raziskovalni pristop znanstvenega raziskovanja (angl. Design Science Research - DSR) v kombinaciji s procesnim modelom CRISP-DM, da se zagotovi metodološka kakovost in praktična relevantnost. Podatki so bili zbrani z avtomatskim izpisovanjem in vključujejo tri modalnosti: metrike profila, besedilne komentarje in vizualne značilnosti. V skladu z Georgesovim modelom identitetnih plasti so bili ti podatki preslikani na deklarativne, aktivne in izračunane vidike digitalne identitete. Zaradi pomanjkanja označenih podatkovnih nizov je bil sprejet delno nadzorovan pristop strojnega učenja, ki izkorišča majhen označen niz skupaj z večjim sklopom neoznačenih računov.

Usposobljenih in ocenjenih je bilo več modelov strojnega učenja, vključno z Random Forest, Support Vector Machines, Logistic Regression in LightGBM. Rezultati kažejo močno učinkovitost modelov, usposobljenih na profilnih metrikah, pri čemer je Random Forest dosegel 97-odstotno natančnost. Nasprotno pa so modeli, ki uporabljajo vizualne značilnosti, dosegli znatno slabše rezultate. Pri uporabi na večjem neoznačenem nizu podatkov je Random Forest označil okoli 2 % računov kot robotske, medtem ko jih je LightGBM identificiral skoraj 18 %, kar odraža različne občutljivosti in predpostavke v modelih. Polnadzorovani pristop je razširil podatkovno zbirko za usposabljanje s 634 na 9562 označenih primerov, pri čemer je 89,3 % neoznačenih podatkov označil z visoko stopnjo zanesljivosti. To dokazuje učinkovitost metode v okoljih z malo označenimi podatki, ki so pogosta v raziskavah na Instagramu.To magistrsko delo prispeva k prikazu potenciala multimodalnih, delno nadzorovanih pristopov za odkrivanje robotov na Instagramu. Poudarja tudi metodološke izzive, kot so omejeni označeni podatki in razložljivost modelov, hkrati pa nakazuje smernice za prihodnje raziskave, vključno z nevronskimi mrežami, analizo sentimenta in preverjanjem obrnjenih slik.

KLJUČNE BESEDE: Družbena omrežja, strojno učenje, delno nadzorovano učenje, raziskovalni pristop znanstvenega oblikovanja, CRISP-DM

CILJI TRAJNOSTNEGA RAZVOJA

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

sl. – Slovene

AI – Artificial Intelligence

AJAX – Asynchronous JavaScript And XML

CRISP-DM – Cross Industry Standard Process for Data Mining

CSS – Cascading Style Sheets

CSV – Comma Separated Values

DIT – Dead Internet Theory

DSR – Design Science Research

HTML – HyperText Markup Language

ID – Identificator

IS – Information Systems

JSON – JavaScript Object Notation

ML – Machine Learning

OSN – Network

PCA – Principal Component Analysis

SNS – Social Network Site

XML – Extensible Markup Language

XPath – XML Path Language

1 INTRODUCTION

The development of Web 2.0 has significantly reshaped the way individuals interact and communicate. It gave rise to social media platforms that empower users not only to consume but also to generate and distribute content and even opening new marketing channels for companies (O'Reilly, 2005, n. p.; Lammenett, 2024, p. 476; Kilian & Langner, 2010, p. 133). Among these platforms, Instagram, along with X (formerly Twitter) and Facebook, plays a central role in digital communication. Instagram offers their users the ability to connect, engage, and build digital identities through interactions such as likes, comments, and shares (Reda, 2024, p. 13). The nature of these interactions is increasingly shaped by underlying recommendation algorithms that curate content and influence user behavior. This is done by observing the engagement for content uploaded and estimating the further likelihood of engagement by other users (Narayanan, 2023, p. 18). In the context of algorithmic influence and large-scale user engagement, social media bots have emerged. Building on these developments, critical theories such as the Dead Internet Theory (DIT) have arisen, suggesting that a substantial share of online activity, particularly on social media, is generated not by humans but by automated bots and artificial intelligence (Muzumdar et al., 2025, pp. 68–69). Wellknown areas of usage for those social bots are: the participation in political discussions to support certain opinions (Hagen et al., 2020, p. 18-19), influencing discussions about the stock market on microblogging platforms like X (Fan et al., 2019, p. 21-22), artificially inflating the engagement with certain users (Muzumdar et al., 2025, p. 70), or being able to influence the three reputation of its own account with systematically requesting users to follow them throughout social networks (Aiello et. al., 2021, p. 16).

The concept of the DIT originated in online forums during the early 2010s and was initially not supported by scientific research (Muzumdar et al., 2025, p. 69). Over time aspects of the theory have attracted attention in academic contexts. Researchers such as Ferrara et al. (2016) examined similar phenomena, including the use of bots to influence political discourse. Since then, interest in detecting synthetic behavior and automated accounts on social media has steadily increased, leading to a broader examination of the implications of such activity (Cresci., 2020). Existing research had addressed the problem of social bot detection using a variety of tools and frameworks. In their comprehensive review, Habib et al. (2024) highlight that a significant portion of scientific literature focuses on social bots on X. Often, studies utilize only a limited subset of publicly available profile data, such as sentiment analysis for detecting fake news in political discussions, image processing techniques for identifying fraudulent accounts, or general engagement metrics. Furthermore, a substantial number of published papers rely on predefined, publicly available datasets with assigned labels, using machine learning (ML) algorithms to classify accounts without additional supervision (Habib et al., 2024, pp. 4–7). In contrast, Goyal et al. (2023) propose a multimodal approach that incorporates graphical content, sentiment analysis, and profile metadata to enhance the detection of synthetic accounts on X (Goyal et al, 2023, pp. 5–7).

The purpose of this master's thesis is to gain a deeper understanding of automated interaction on social media platforms originating from the influence of algorithmic behavior and artificial intelligence. To reach the purpose, this thesis follows the following goals:

- To investigate bot-behavior on social media and distinguish between types and use cases of them.
- To examine approaches for distinguishing between authentic and inauthentic profiles using data-driven classification techniques.
- To implement and evaluate a machine learning-based model for detecting fake Instagram accounts.
- To analyze the results in light of existing research and contribute to the ongoing discourse on synthetic behavior in social media environments.

Several types of methods and techniques are going to be used while developing the master thesis. In the theoretical part, a descriptive and analytical method will be used to provide a structured background for the research. Existing academic literature on social media, fake engagement, machine learning (ML), and the DIT will be reviewed. The literature will include foundational previous findings on bot behavior, and social media. Through this, a conceptual framework will be developed to support the empirical part of the thesis. The empirical part of the thesis will be based on the development and evaluation of a ML model designed to detect fake profiles on social media platforms, with a particular focus on Instagram. Freely accessible public data will be collected through web scraping techniques and API's including user comments, profile metadata, and engagement metrics. From this data, relevant features will be extracted, such as account age, follower, amount timing of interactions, and characteristics of comments. These features will be used as inputs for ML algorithms such as Random Forest, LightGBM, or Logistic Regression.

The development of the ML model and its application for bot detection will be structured according to the 'Cross Industry Standard Process for Data Mining' (CRISP-DM) framework. This will ensure a systematic and standardized approach to the data mining process (Wirth & Hipp, 2000, p. 1-2) through its predefined phases including; business understanding, data understanding, data preparation, modeling, evaluation and deployment (Wirth & Hipp, 2000, p. 5-7). In parallel, the overall structure of the thesis will follow the Design Science Research (DSR) methodology, which is suitable for research focused on creating and evaluating artifacts that solve real-world problems (Hevner et. al., 2004, p. 79). DSR will guide the identification of the problem, the design and implementation of the artifact, in this case the ML model, and its evaluation and deployment. Particular attention will be paid to ethical considerations; all collected data will be anonymized where necessary, and no attempts will be made to trace information back to individual users. By utilizing supervised and semi-supervised classification techniques, the thesis will contribute to a better understanding of synthetic behavior in social media environments. This research seeks to build on and extend the growing body of literature on digital authenticity and bot detection while providing a structured approach to analyzing the claims suggested by the DIT.

The structure of the thesis reflects this orientation. Chapter 2 presents the theoretical background, covering social media dynamics, types and use cases of bots, and prior research on ML for bot detection. Chapter 3 explains the methodological framework, combining DSR with the CRISP-DM process to ensure both rigor and practical relevance. Chapter 4 contains the practical implementation, where the dataset is introduced, prepared, and modeled. Here, both baseline models and semi-supervised approaches are evaluated. Chapter 5 discusses the results in relation to existing literature, identifies limitations, and suggests avenues for future research. Chapter 6 concludes the thesis by summarizing the main contributions and answering the research questions.

2 THEORETICAL BACKGROUND

This particular chapter introduces the main concepts linked to the overall topics. First it is important to understand social media and Online Social Networks (OSN). The most significant platforms will be introduced as well as their platform dynamics and mechanisms to leverage the platform users engagement with the content shared. Afterwards robots and automated behaviour will be introduced. Combining these topics an eye will be layed on the existing literature in detecting these robots on the internet and particullary on social media where several papers will be introduced. The introduction to ML is necessary, as it is the tool of choice used in this thesis to detect synthetic behaviour online.

2.1 Social Media Definitions

The internet has gone through various development over time. At first Hypertext Markup Language (HTML) documents were published which could be observed by online users (Blumauer & Pellegrini, 2008, p. 12). Tim Berners- Lee, who initially developed the world wide web on his own web server, saw large potential in the technology, especially in revolutionizing the communication between individuals. He as well noted that to that time few people had access to so called “hypertext creation tools” which implies that only people which such tools were able to share content (Berners- Lee, 1998). In a conference session where O’Reilly and Media live participated “Web 2.0” came up as a definition for a larger shift in development of the web known until then (Choudhury, 2014, p. 8097). Later on, Tim O’Rilley defines several characteristics which describe this transformation including a more service dominant architecture of services, the utilization of knowledge users can contribute online and the trust of giving users the right to co-create online (O’Reilly, 2005, p. 5). The nature of these developments in the web structure was the fundamental brick making the implementation of social elements in the web more valuable. While Web 2.0 was no update to a newer version of the web it was more an enrichment of features. Some of those features were for example the Adobe Flash enabling visual content as videos or the Asynchronous JavaScript (AJAX) and the Extensible Markup Language (XML) technique which made it possible to refresh elements of a web page without the urge to reload the whole web page.

Based on this progress aligned with the creation of 'User Generated Content' online services arose which could be classified as social media applications (Sykora, 2017, p. 5; Kaplan & Haenlein, 2010, pp. 60-61). The concept of social media has been defined in various ways across academic literature. For the purpose of this thesis, Carr and Hayes (2015) definition was adopted, claimingthat »Social media are Internet-based channels that allow users to opportunistically interact and selectively self-present, either in real-time or asynchronously, with both broad and narrow audiences who derive value from user-generated content and the perception of interaction with others.«

Social media can be used as an umbrella term covering many subgenres including Content Communities like Youtube or Social Network Sites (SNS) like Instagram (Kaplan & Haenlein, 2009, p. 565). While many academic sources do not distinguish between the terminology of social media and SNS this work shall be supported by the definition of SNS as the scope lies particularly on this kind of services. Boyd & Ellison define SNS as web-based services where an individium can establish a profile, close connections with others and observe their own as well as connections of others. Oftentimes those services serve as an extension to existing real world social structures that transfer into digital ones (Boyd & Ellison, 2010, p. 211).

2.2 Social Media Elements and Characteristics

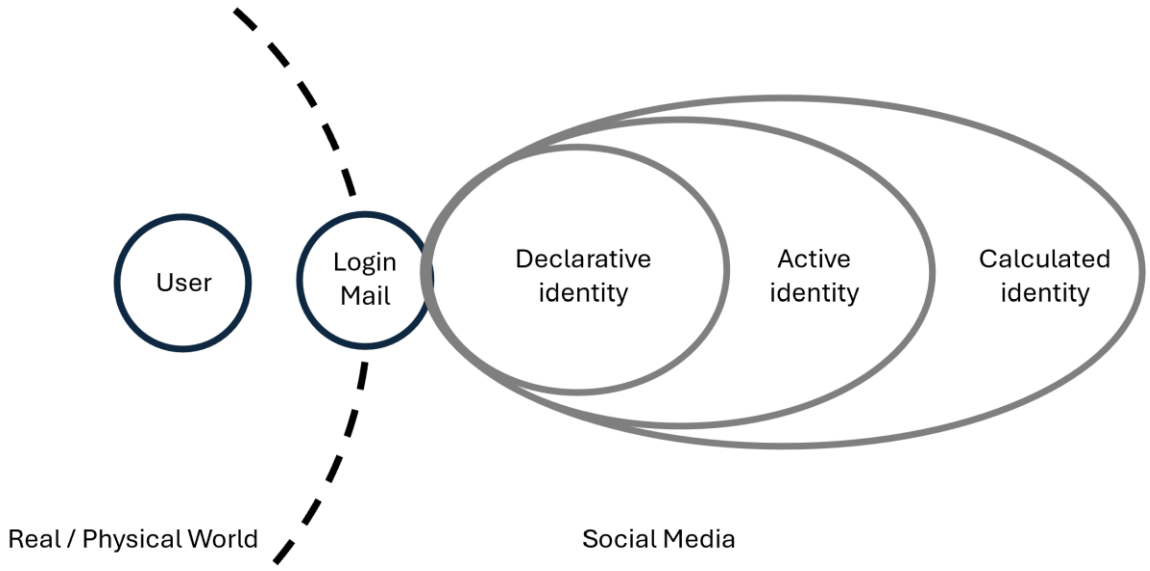
Arising from this it is possible to derive classical social media elements which occur almost anytime at each platform. The first element is the profile element which can include several subfeatures as self-expression in form of media, short introductions or further personalizations. Another key aspect is the connection between users in either one directional manner as following a profile or bidirectional by creating for example online friendships (Bayer et. al., 2020, pp. 476-479). Furthermore the enablement for users to see the connections they have made and observe other accounts for their connections if permitted is another element which characterizes social media (Boyd & Ellison, 2010, p. 213). Lastly the message element became a substantial part of SNS platforms by offering communication between two or more users (Bayer et. al., 2020, p. 483).

The introduction of SNS services rapidly changed many circumstances in human social areas. First, the above-mentioned element of communication significantly influenced the opportunities available to an individual for communicating with others, despite the greater physical distances between them. It revolutionized the speed at which individuals can talk to each other. Notable developments have emerged in how online services have evolved into platforms that offer users the opportunity to comment, review, or rate products, as seen on well-known websites like Amazon.com.

The ability to utilize messaging elements as introduced above led to several tendencies related to communication, such as shortening the number of characters used to express oneself. This doesn't necessarily mean that the value of information decreases, as

abbreviations were often used to convey certain information more shortly, for example, the acronym ASAP, which, when written out, means: as soon as possible (Subramanian, 2017, p. 72). Another aspect that changed dramatically is the presence of users online, as SNS convinced them to create a digital identity, which includes several web elements, such as the already presented profile element. This enables self-representation and the opportunity to convince others of oneself the way one would like to (Reda, 2024, p. 5). Digital identity can be understood as a combination of three parts: the data left behind on digital platforms, how individuals use others' content to shape how they appear online, and the personal image they form of themselves through their interactions with technology (Georges, 2011, p. 33). This identity can then be split down into specific elements, which can be seen in Figure 1.

Figure 1: Georges identity layers



Source: adepted from Georges (2011).

The declarative identity is the one with the highest share of control exerted by the users themselves. It consists of the statements one declares by creating and curating the profile. Active identity is the trace a user leaves by interacting with others, where others, for example, are notified that another user commented on their new post. The calculated identity refers to the metrics underlying a user's network, such as the number of followers on a specific account (Georges, 2011, p. 39). These identity layers, once embedded in SNS, became more than personal representations; they evolved into data points that drive interaction, visibility, and even economic value (Van Dijck, 2013, p. 4).

2.3 Success Factors of Online Social Networks

As social media platforms gained traction, their success was not solely rooted in technological affordances but also in the social structures they enabled and amplified.

Central to this expansion was the phenomenon of network effects, a concept that helps explain how user adoption and platform value are intrinsically linked. Network effects suggest that the utility of a service increases with the number of its users. In the context of SNS, each new user contributes content, establishes connections, and potentially engages with others, thereby enriching the network and reinforcing participation among existing users (Sander & Teh, 2014, p. 267). Metcalfe's Law, for example, suggests that a network's overall value increases exponentially in relation to its user base, meaning that as more individuals join, the potential connections and the utility grow significantly (Metcalfe, 2013, p. 26). Despite criticism towards Metcalfe's law, a study by Zhang, Liu, and Xu provided empirical validation using real-world data from Facebook and Tencent. Their analysis showed that among various models describing network value, Metcalfe's quadratic relationship between network size and value best fit the data, reinforcing its relevance for describing the growth and economic potential of social networks (Zhang et al., 2015).

A further interesting factor contributing to the success of the internet and social media lies in the emergence of what Castells describes as mass self-communication. This form of communication allows individuals to create, share, and consume content on a global scale, independently of traditional media structures. Its decentralized nature, combined with digital accessibility, enables widespread participation and reinforces social media's role as a platform for personal expression and cultural production. Especially the traditional era of content distribution dominated by media institutions has been complemented by many-to-many communication, where individuals can access information from a wide range of actors on social media platforms (Castells, 2010, p. 31).

Lastly, an additional reason for the success of SNS lies in the existence of connections and engagement. A connection can be viewed as an existing online friendship, where engagement is demonstrated through likes or comments online. Successful platforms have recognized that social interaction is a crucial aspect in driving platform usage and, consequently, retention (Schultz et al., 2014, p. 1). Therefore, SNSs like Facebook design the platform to encourage friendships not for their intrinsic value but to drive user engagement. This is reflected in the platform's emphasis on visible metrics as likes and comments, which not only quantify social connections but also shape how users perceive and interact within the network (Grosser, 2014, p. 3; Bucher, 2018, pp. 11-12). Building on this, Di Gangi and Wasko demonstrate that user engagement, shaped by both social factors such as the presence of a critical mass of acquaintances and technical features like platform evolvability, plays a crucial role in driving continued social media usage. Their Social Media Engagement Theory highlights that platforms are intentionally designed to foster not just interaction but sustained psychological involvement (Di Gangi & Wasko, 2016, pp. 10-13).

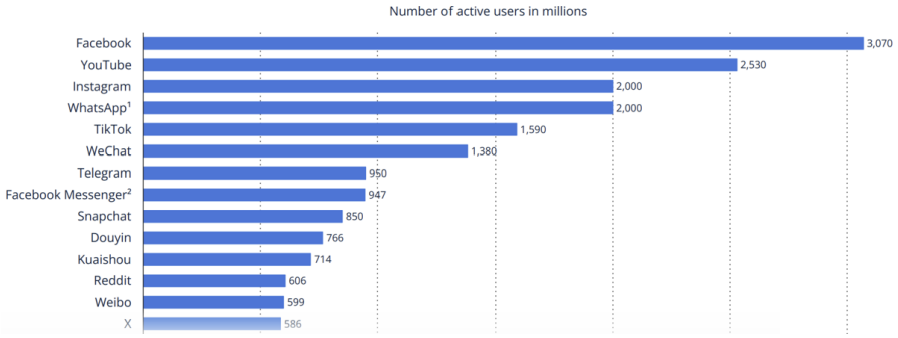
Similarly, Ray et al. show that in online communities, engagement goes beyond user satisfaction and emerges as a central force behind active participation and content contribution. When users feel emotionally connected and perceive their involvement as meaningful, they are more likely to stay engaged and promote the platform (Ray et al., 2014, pp. 542-544). Together, these findings reinforce that engagement is not a byproduct of platform use: it is a strategic objective that underpins user retention and activity in SNS.

In conclusion, the growth of SNS has been propelled by a combination of technological affordances and socially driven mechanisms like network effects, participatory communication, and engagement-focused design. These elements have not only fueled rapid adoption but also shaped the ways users interact and remain active on platforms. However, while these dynamics explain the initial rise of social media, they do not fully account for why some platforms have endured while others have faded. To understand this persistence, the next chapter offers a brief timeline of dominant platforms such as Facebook, X, Instagram, and TikTok. Additionally, the chapter explores how these platforms differ in the roles they serve, whether visually oriented, text-based, or short video-focused, and how these differences shape distinct models of user interaction across the social media landscape.

2.4 Dominant Social Network Sites

Since the evolving era of the web, as mentioned in the preceding chapters, several SNS have been brought to life. Since SNS became popular in the middle of the first decade of the current century, the usage of social media platforms has constantly increased. 2.73 billion people were using social media back in 2017, while this number increased to 5.17 billion in 2024 and is expected to be twice as large compared to 2017 in 2025. Globally, social media reached over two-thirds of the population according to the social network penetration rate (Statista, 2024, pp. 4-5). The top 3 leading social networks according to Statista are Facebook with a total of 3.07 billion active users, followed by YouTube with a total of 2.53 billion active users, as displayed in Figure 2. Instagram, in third place, has a total of 2.00 billion active users worldwide (Statista, 2024, p. 12).

Figure 2: Most popular social networks in 2025, by monthly active users in millions



Source: Statista (2024, p. 12).

Facebook was initially launched in 2006 as a SNS exclusively for students in the United States. Over time, it expanded its accessibility to a broader audience, including private individuals as well as businesses (Weinberg & Pahrmann, 2014, p. 223). The platform distinguished itself from other services for several reasons, particularly through features designed to establish trust and ensure user privacy. Among these was the implementation of customizable privacy settings, which allowed users to control the visibility of their content by restricting it to specific groups of friends. Another factor contributing to Facebook's early success was its strategy of organic growth, which involved gradually expanding access by including additional academic institutions in a phased manner (Shih, 2011, pp. 20–21). Furthermore, Facebook successfully fostered user engagement by introducing features such as the news feed and interactive content functionalities. These included birthday reminders for connections and the integration of third-party services, such as games, which further enhanced user retention and platform interaction (Shih, 2011, p. 21; van Dijck, 2013, p. 71; Weinberg & Pahrmann, 2014, p. 226).

YouTube was launched in 2005, prior to Facebook, enabling users to upload video-based content intuitively and reach a global audience. The audience can observe these videos free of charge on so-called »channels« which platform visitors can engage with through subscription (Burgess & Green, 2010, p. 1; McDonough et al., 2021, p. 155; Taylor et al., 2011, p. 259). While YouTube is sometimes mistakenly associated with early peer-to-peer file-sharing platforms, its core innovation lay not in new transmission technologies but in combining video streaming, user-generated uploads, and social features. These elements, operating within a centralized infrastructure, marked a shift from traditional broadcasting by enabling participatory and on-demand media consumption (Van Dijck, 2013, p. 112).

Instagram, launched in 2010, quickly established itself as a leading social media platform by emphasizing visual content and a mobile-first, user-friendly interface. In contrast to Facebook, which originated within academic networks and gradually expanded into broader social functionalities, Instagram was designed from the outset as a platform dedicated to the sharing of aesthetically curated photos and, later, videos (Hu et al., 2014, p. 596). In 2012, Facebook acquired Instagram for approximately one billion US dollars. According to Facebook CEO Mark Zuckerberg, the intention behind the acquisition was not to integrate Instagram's functionalities into Facebook, but rather to support Instagram's independent growth, increase its global visibility, and allow users to choose between both services freely (Facebook, 2012). At its core, Instagram initially succeeded by integrating various features already familiar to users from other platforms, most notably the ability to upload photos and enhance them using stylistic filters, thereby aligning content with prevailing visual trends and user expectations (Leaver et al., 2020, pp. 21–22). Over time, Instagram introduced a number of additional functionalities aimed at increasing user engagement and interaction. These included hashtags, which allowed content to be organized and discovered through topic-based clustering, and the "Stories" feature, enabling users to share temporary content visible for 24 hours (Chen et al., 2021, pp. 203–214; Hu et al., 2014, p. 596).

After outlining the role of dominant SNS, it becomes clear that their success is closely tied to the way content is presented to users. Central to this process are recommendation algorithms, which shape visibility and engagement.

2.5 Recommendation Algorithms

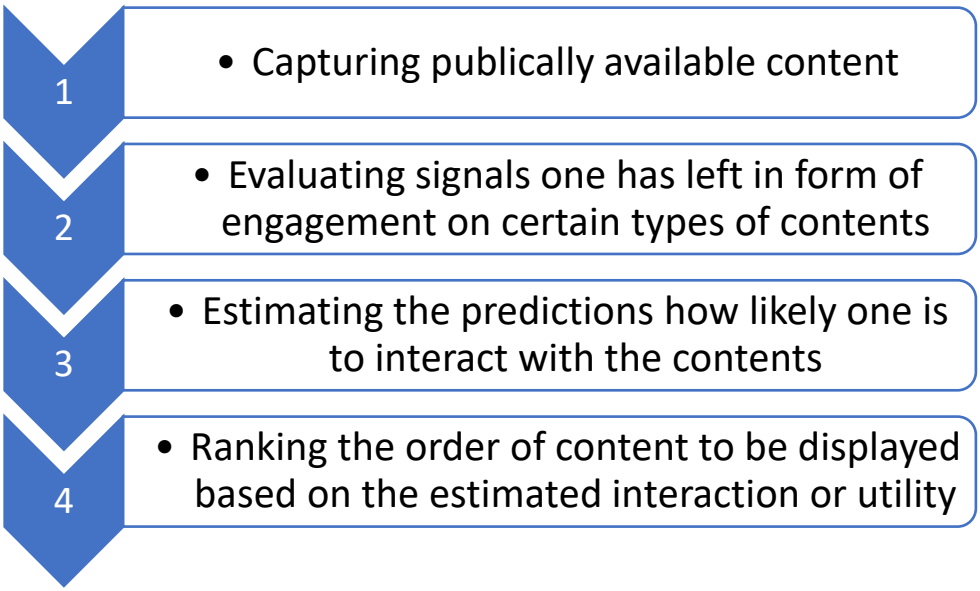
From the early beginnings of social media, platforms followed a particular logic in how they presented content to users. Many services, including Facebook and Instagram, initially displayed content in the form of a chronological feed. The feed was initially shaped by individual choices, based on which accounts or channels a user decided to follow and receive updates from (Narayanan, 2023, p. 10). YouTube, on the other hand, prioritized content that gained popularity quickly. As a result, many users were shown the same videos at the same time when visiting the site, regardless of their individual preferences (Goodrow, 2021). Facebook later identified that purely chronological feeds contributed to an overwhelming volume of content, especially for users who had been inactive for a longer period. In such cases, users were more likely to disengage when confronted with a backlog of time-sensitive content (Meta for Business, n.d.). The growing number of users on these platforms further exacerbated the problem, as more connections led to more shared content (Statista, 2024, p. 14). To address this issue, recommendation systems were introduced (Narayanan, 2023, p. 40). These systems rely on algorithmic logic to optimize the user experience by predicting which content is most likely to lead to interaction or engagement. At their core, recommendation algorithms aim to personalize content delivery by analyzing large volumes of data, often referred to as "signals," which vary depending on the platform (Narayanan, 2023, p. 22; Shokeen & Rana, 2019, p. 966).

According to Narayanan, these signals can be categorized into several domains. The first domain is the user network, which includes indicators such as social connections and interactions, for example, when users like or comment on each other's content. The second domain is user behavior, which refers to how individuals engage with specific topics or types of posts on the platform. The third domain comprises demographic data, including factors such as age, gender, and geographic location. The objective of a recommendation system is to match users with similar signals and, based on this, to estimate the probability that a given user will engage with a particular post. Similarly, the system can also assess the likelihood that specific content will attract attention from predefined user segments (Narayanan, 2023, pp. 22–24).

From a broader perspective, the overarching goal of recommendation systems is to ensure that users are continuously presented with content that aligns with their interests (Shokeen & Rana, 2019, p. 966). By maximizing user satisfaction through relevant content, platforms can increase engagement, which is a key metric for business objectives such as user retention and advertising revenue (Bucher, 2018, p. 6; Narayanan, 2023, p. 18). Platforms have recognized the strategic value of such systems and implemented them accordingly. Among

the first to adopt algorithmic recommendations was YouTube, which introduced a hybrid system as early as 2008. This model combined content from a user's subscriptions with personalized video suggestions based on algorithmic predictions. YouTube's signals for content ranking include user clicks, viewing duration, likes, shares, and comments, which were then processed by the algorithm to generate further recommendations (Goodrow, 2021). Facebook followed with the implementation of recommendation systems in the news feed, additionally supported with a 'Ticker'-feed where all information is chronologically ordered by 2011 (Bucher, 2018, p. 74). Facebook stated that the immense flow of unfiltered information would lead to decreasing engagement. The signals utilized for the recommendations range from the number of visits one takes to different Facebook pages to the number of times one interacts with a specific category of posts, and of course, likes and comments (Backstrom, 2013). Meta, as the mother company of Facebook and Instagram (Meta, n. d.), published a simplified procedure on how the recommendations are generated for Facebook and Instagram through their transparency center:

Figure 3: How recommendations are generated on Meta Platforms



Source: Meta (2025b); Meta (2025c).

As can be seen in Figure 3, the aim is to maximize the engagement on the platform by utilizing the previously captured signals. Instagram, as it is part of Meta, works in a similar way as Facebook (Meta, 2025b; Meta, 2025c). The signals used by Instagram are not different from those Facebook uses. Instagram uses metadata of posts, interaction patterns between users, likes, comments, and the content one saves in its archive (Mosseri, 2021).

In summary, the evolution from chronological feeds to algorithmic recommendation systems has fundamentally transformed how content is distributed and consumed on social media platforms. What began as user-driven information streams has shifted towards predictive models that rely heavily on interaction-based signals. Likes and comments play a central

role as they are a primary input for the algorithmic logic that determines content visibility. The capacity of recommendation systems to learn from such feedback and continuously adapt has made them indispensable for platform operators striving to maximize user retention and advertising revenue. Engagement, in this context, is not merely a measure of success but a mechanism that shapes the very architecture of content delivery. This development raises important questions about the nature and authenticity of interaction on these platforms. As the next chapter will explore, concerns have been growing around the synthetic manipulation of engagement metrics, particularly through the rise of non-human activity.

2.6 The dead internet theory

The rapid advancement of Artificial Intelligence (AI) has brought fundamental changes to how online content is generated and distributed. On SNS, algorithmic recommendation systems increasingly prioritize content that is most likely to generate user interaction. As a consequence, the authenticity of digital content and engagement has come under growing scrutiny (Walter, 2024). This skepticism is obvious on platforms such as X, where users have repeatedly raised concerns about posts that appear low-effort or unremarkable but attract disproportionately high levels of interaction. Such dynamics have led to increasing mistrust toward the integrity of social signals, particularly when coordinated campaigns or automated engagement patterns are suspected (Di Placido, 2024). Over the years, recurring doubts about the authenticity of online interactions have been observed, particularly in relation to coordinated campaigns in the gaming sector or politically motivated bot networks (Mariani, 2023, p. 35). From such speculations, a broader narrative emerged in various online forums, commonly referred to as the DIT. This theory, often characterized as conspiratorial in nature, suggests that the internet as a space for genuine human interaction effectively "died" around a decade ago (Sheehan, 2021).

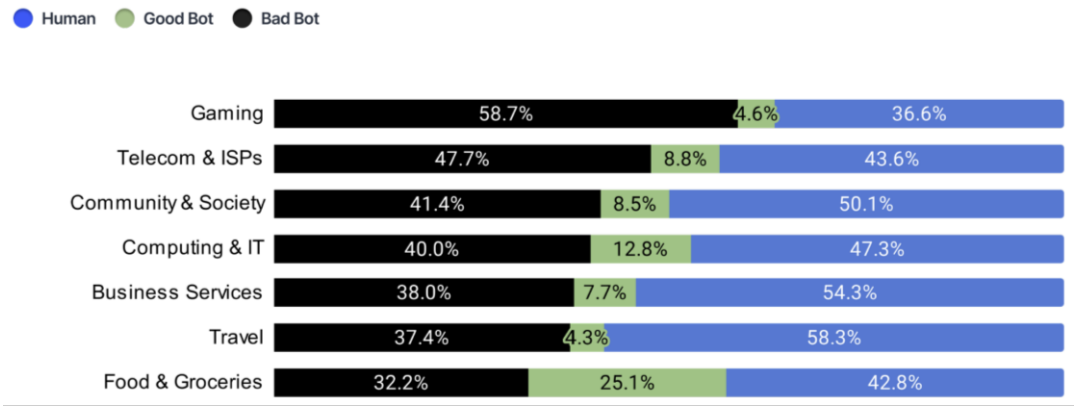
While the precise origin of the DIT cannot be clearly identified, it is widely believed to have developed across various online forums such as ‘4chan’, where it was discussed in fragmented form. The theory gained broader attention following the publication of a detailed post on the platform ‘Agora Road’, which sought to substantiate its claims and convince a wider audience of its validity (Di Placido, 2024; Sheehan, 2021). Although speculative in nature, the narrative began to take shape toward the end of the 2010s and has since gained traction in both online discourse and academic reflections on digital authenticity (Di Placido, 2024; Muzumdar et al., 2025, p. 69). The central premise of the DIT suggests that the internet, once characterized by organic user interaction, has increasingly been overtaken by synthetic activity. As a result, genuine human engagement is perceived to have become marginal, if not rare, within many online environments (Muzumdar et al., 2025, p. 68; Walter, 2024). A further core aspect is that most online users are confronted with the inability to tell the difference between real and synthetic interaction (Arjel et al., 2025, p. 116). Many aspects arise from the presented claims, including:

-An aspect that supports the assumptions of the DIT is the use of recommendation systems implemented by major social media platforms such as Instagram and YouTube. The strategic deployment of content farming within algorithmically curated environments has led to an increasingly homogenized digital landscape, where content visibility is primarily determined by its engagement potential rather than its quality or originality. This mechanism not only enhances platform profitability but also reduces opportunities for authentic interaction and user-generated contributions, in favor of algorithm-optimized, attention-grabbing material (Muzumdar et al., 2025, p. 70).

- Content creation and distribution have undergone a rapid transformation, with AI enabling the fast and large-scale generation of various media formats (Arjel et al., 2025, p. 117). The distinction between AI-generated and human-produced content is becoming increasingly difficult to discern, as generative models continue to improve in realism and complexity (Walter, 2024). This is particularly concerning in the context of political discourse and news media, where technologies such as deepfakes pose challenges to content authenticity and moderation. Moreover, AI has significantly improved the capabilities of bots, enabling them to participate in online discussions with increasing coherence and contextual awareness (Arjel et al., 2025, pp. 117–118). As a consequence, algorithmically optimized content is becoming more dominant, gradually displacing human-created material and raising critical concerns about the integrity, diversity, and informational value of today's online platforms (Muzumdar et al., 2025, p. 70).

- Lastly, the rise of algorithmic behavior in the form of bots represents a core element that supports the central assumptions of the DIT. These bots are typically programmed to simulate human activity and to perform specific tasks within digital environments (Muzumdar et al., 2025, p. 70). The cybersecurity company Imperva, which provides a wide range of security solutions, has conducted investigations into this phenomenon (Imperva, 2023, p. 44). Their findings indicate that in 2022, 47.4% of total online traffic was generated by bots, while only 52.6% could be attributed to human users. A more detailed analysis by industry revealed that the category "Community and Society", including social media platforms, was among the most affected sectors, with 49.9% of its traffic attributed to bots, placing it among the top industries in terms of bot activity, as it can be seen in Figure 4 (Imperva, 2023, pp. 11–12).

Figure 4: Share of Bots against Himan broken down by industries



Source: Imperva (2023, p. 11).

Altogether, the increasing prevalence of AI-generated content, algorithmically curated engagement, and automated activity highlights a growing tension between perceived authenticity and actual interaction within digital spaces.

2.7 Social Bots

As shown in the previous chapter, bots are a key part of the DIT and are especially important in social media settings. This thesis focuses on the role of bots as the most straightforward part of the theory to study. While algorithmic recommendation systems certainly help create synthetic engagement, they mainly work as "black boxes," making it hard to analyze how they function without direct access to the platform (Narayanan, 2023, p. 5). Similarly, AI, although closely related to content creation and automation, mainly acts as a tool that enables bots (Cresci, 2020, p. 77). Therefore, this research highlights the detection and analysis of social media accounts that behave algorithmically to tell apart real human users from accounts that may be operated or boosted by bots. The next chapter will introduce key definitions and theoretical basics related to the use of bots, especially their role within social media environments.

2.7.1 Definition

A bot is a software program that works independently without human control. The algorithm behind a bot is usually created to achieve a specific goal, which can include tasks that are simple and repetitive or more complex functions such as engaging with users or affecting the visibility of digital content, especially in the case of social media bots (Imperva, 2023, p. 4; Carley & Ng, 2025, p. 3; Imperva, n.d.).

Social media bots, or social bots, have multiple definitions, with no single, universally accepted one in the academic literature. To fill this gap, Ng and Carley proposed a comprehensive definition, stating that a social bot is "an automated account that carries out a series of mechanics on social media platforms, for content creation, distribution, and collection, and/or for relationship formation and dissolution" (Carley & Ng, 2025, p. 4).

Social bots serve a wide range of purposes. As the Imperva report (2023) notes, bots are often classified as either beneficial or harmful, based on their intended use. Non-malicious uses include situations where celebrities or online communities employ bots to communicate more efficiently with large audiences (Ng et al., 2016, p. 9). Likewise, companies increasingly depend on social bots, like chatbots, to support customer service on social media, which helps reduce workload and speeds up responses (Aljabri et al., 2023, p. 19; Ferrara et al., 2016, p. 96).

2.7.2 Use cases

However, not all social media bots serve benign purposes. The same automation features can be exploited to support manipulative or deceptive goals. A notable example is the use of political bots, which are deployed to interfere in political discourse, amplify partisan viewpoints, or flood discussions with coordinated narratives. These bots may spread low-credibility or false content, mention influential political figures to simulate legitimacy, and create artificial consensus through mass-posting, especially on platforms like X (Kollanyi et al., 2016, p. 1; Shao et al., 2018, p. 2). During the 2016 U.S. Presidential election, researchers found that approximately 400.000 bots were responsible for generating around 3,8 million tweets, accounting for a large part of the political conversation. These bots played a key role in spreading misinformation, increasing polarization, and boosting the influence of inauthentic actors within political debates (Bessi & Ferrara, 2016, p. 11).

Social bots have also been seen in financial areas, especially in changing how the public views stock market trends. They often do this by posting large amounts of speculative content to sway investor actions (Ferrara et al., 2016, p. 96). Talavera and Tran (2019) discovered that, while bot-created tweets did not have a statistically meaningful impact on stock returns, the sentiment of these tweets was significantly linked to stock volatility and trading volume. These results imply that social media activity can affect financial market behavior, particularly regarding perceived uncertainty and trading patterns (Talavera & Tran, 2019, pp. 21–22).

Another use case involves engagement manipulation, where bots artificially boost the visibility of posts or accounts. This includes purchasing likes, shares, or followers to give the impression of popularity or influence. A well-documented incident occurred in May 2020, when a fake Weibo post claiming abuse by a schoolteacher went viral. The post gained significant traction through bought fake engagement, ultimately sparking public outrage and

an official investigation. It was later revealed that the content had been intentionally fabricated and strategically amplified using social bots (Fang & Nie, 2023, p. 3272).

2.7.3 Behaviours

The latest advancements in AI enhance the development of social bots and their ability to conceal their artificial nature (Radivojevic et al., 2024, pp. 6-7). To understand the behavior of these bots, it is essential to identify the specific goal they were programmed to achieve. In this thesis, the focus is on the synthetic engagement of these entities to artificially boost posts and inflate the popularity of accounts and posts online.

As previously discussed in the context of social media environments, user identity can be conceptually divided into three dimensions: declarative, active, and calculated identity (Georges, 2011). This framework not only explains how real users present themselves online but also provides a useful perspective for understanding and identifying non-human actors whose synthetic behavior can often be detected through inconsistencies or anomalies across these identity layers. In this chapter, the framework is revisited to analyze how social bots differ from typical human behavior and how such deviations can inform feature selection and classification processes in this thesis.

The declarative identity includes information a network user shares about themselves, such as uploading a profile picture, listing personal interests, or posting content (Georges, 2011, p. 40). Signs of suspicious behavior in this area often include the lack of common features typically found on genuine user profiles, like a profile picture or a bio (PWC, 2017, pp. 11–12). Bots, unlike human-operated accounts, often hide or omit identity signals (Carley & Ng, 2025, p. 6). Studies have shown that bots usually use generic or randomly generated usernames, often made up of nonsensical combinations of letters and numbers. They also tend to use default or duplicated profile images, some of which are taken from other users and reused across multiple bot accounts (Beskow & Carley, 2018, p. 4).

The active identity refers to the behavior users exhibit through their interactions, such as liking content, commenting, following, or initiating contact with others (Georges, 2011, pp. 40–41). In this context, bots often display unnatural patterns, including excessive engagement like mass-liking posts or repeatedly commenting and sparking through rapid reactions to newly published content. These accounts may also show abnormally high posting frequency or repetitive comment content that often lacks contextual relevance (PWC, 2017, pp. 11–12). Bots are also known to send large numbers of friend or follow requests, which sets them apart from human users. For example, Wang et al. (2013) report that bots spend approximately 41% of their activity on sending friend requests, compared to just 0.3% for genuine users. Additionally, their behavior tends to follow time patterns, operating in short, intense bursts aligned with pre-programmed schedules (Carley & Ng, 2025, pp. 6–7).

Finally, the calculated identity includes metrics generated by the platform, such as the number of followers, followings, or posts an account has made (Georges, 2011, pp. 40–41). Bot accounts generally show limited social connectivity, often being part of artificial networks made up of other bots (Goyal et al., 2023, p. 2). This limited and isolated network of connections makes them easier to detect, as genuine users can more readily identify suspicious accounts based on unfamiliar or irrelevant friend suggestions (PWC, 2017, pp. 11–12).

2.7.4 Mitigation

Meta uses various strategies to detect and prevent synthetic activity on its platforms. First, the company depends on user-reported reports, where real users can flag suspicious or inappropriate behavior that may suggest automated accounts. Additionally, Meta imposes rate limits on certain activities, like content uploads and interactions, to reduce spam and identify unusual usage patterns. When these behavioral thresholds are exceeded, automated systems are activated to flag and report the accounts involved (Meta Platforms Ireland Limited, 2022, pp. 2–4). To address identity theft and the misuse of celebrity images, Meta has adopted facial recognition technologies that can identify famous individuals and flag potentially fraudulent use of their pictures (Meta, 2025d).

However, these systems have limitations. Bots that use default profile pictures, avoid high activity levels, and post non-controversial content often go unnoticed. Additionally, such accounts are unlikely to be reported by users if they do not cause visible harm or disruption.

As Meta itself acknowledges, detecting increasingly sophisticated social bots is becoming more difficult. In response, the platform has provided access to specific data via Application Programming Interfaces (API), enabling researchers to investigate synthetic behavior patterns (Meta Platforms Ireland Limited, 2022, pp. 26–27).

The following chapter reviews existing literature on bot detection, tracing its development from early research to more recent, platform-specific methods. Special focus will be on classifying detection techniques by social media platform and examining the features used to distinguish genuine from fake user behavior.

2.8 Existing Literature

This chapter provides a comprehensive review of existing research approaches to social bot detection, examining both traditional non-ML methods and contemporary ML applications.

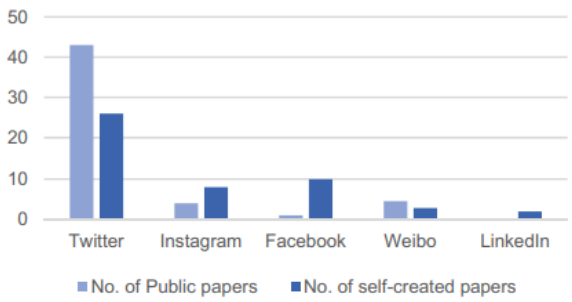
2.8.1 Early approaches

The early 2010s marked the beginning of research into bot detection. Various methods emerged, each with a different approach to observation or identification. Wang et al. (2012) suggested using crowdsourcing as a detection technique, where multiple people are asked to distinguish between real and synthetic users online. Other efforts involved using so-called honeypots, which serve as bot traps. The researchers created accounts on the platform X, posting nonsensical content to attract spammers and malware, and collected 36.000 accounts that responded (Lee et al., 2011, p. 1). Later, researchers found that many of these responses could be classified as social bots (Ferrera et al., 2016, p. 99).

2.8.2 Platform-centric approaches

While these early methods laid the foundation for bot detection, their effectiveness varied greatly across different platforms. Many studies were focused on platform X (shown as Twitter in Figure 5). This likely stems from the platform's open nature, which allows public access to extract user data from both public and private profiles. Not only is the availability of data appealing, but the sheer volume of information that can be pulled is also significant; for example, Shao et al.'s research on X involves millions of publications (Beskow & Carley, 2018, p. 4; Shao et al., 2018, p. 6). Habib et al. examined the diversity of bot detection research that uses only publicly accessible data. They found that 5 out of 8 studies on fake profiles were conducted solely on X, and 4 out of 5 studies on social bot detection used X as the social media platform (Habib et al., 2024, pp. 4-5).

Figure 5: Distribution of research papers on social bots



Source: Aljabri et al. (2023, p. 19).

The availability of large public data sets has enabled researchers to achieve impressive detection accuracy, with reported rates reaching 97.0% (Sengar et al., 2020, p. 43) and even 99.4% in some studies (Carley & Ng, 2025, p. 5). While X has dominated this research area (as shown in Figure 5), significant work has also been done on other platforms. Santia et al. (2019) showed that although bots made up only 0.06% of Facebook users, they accounted for about 10% of comments on news posts. Similarly, Dewan and Kumaraguru built a real-time bot detection system analyzing over 4 million Facebook posts, reaching 86.9% accuracy

with a browser plugin and public API implementation Dewan and Kumaraguru, 2015, pp. 9-10).

Research on Instagram remains relatively limited (Aljabri et al., 2023, p. 20), but notable contributions include Zarei et al.'s (2019) study of impersonator accounts within political and sports communities. Later research has identified key Instagram bot indicators based on profile features, such as follower-following ratios, username patterns, and verification status (Akyon & Kalfaoglu, 2019, pp. 3-4; Meshram et al., 2021, p. 121).

2.8.3 Feature-based approaches

Beyond platform-specific analyses, research on social bot detection has increasingly utilized feature-based approaches. These methods analyze behavioral patterns and profile characteristics that differentiate automated accounts from human ones (Ferrara et al., 2016, p. 101). As observable in Appendix 1, Ferrara et al. created a detailed classification system that organizes these features into specific categories, which later studies have further validated through empirical testing.

2.8.4 Limitations

The evolution of social bots has made traditional detection methods less effective, as the behavioral lines between automated and human accounts continue to blur (Cresci, 2020, pp. 77–78). This issue is demonstrated by the limited success of crowdsourcing methods, which correctly identified only 24% of bots in a notable study (Cresci et al., 2017, p. 6). The mid-2010s marked a key shift as the usage of supervised and unsupervised ML algorithms increased in bot detection research (Cresci, 2020, p. 78).

The following chapter will explain how ML can fill these gaps by integrating and analyzing various behavioral features at the same time. This method provides the foundation for the practical implementation discussed in later sections of this thesis.

2.9 Machine learning

Despite methods like crowdsourcing, which rely heavily on human judgment, or rule-based approaches where certain thresholds determine classifications, ML independently extracts patterns from data (Géron, 2019, p. 4). This approach has proven especially useful in data-heavy fields where traditional techniques, such as manual observation, are limited in scalability or efficiency (Murphy, 2012, p. 1). By analyzing structured input data, ML enables pattern recognition, predictive modeling, and the discovery of new insights, supporting human decision-making across a variety of uses (Kuhn & Johnson, 2018, p.1; Mitchell, 1997, pp. 1-2; Murphy, 2012, p. 1). These applications range from simple tasks like spam detection to complex forecasting such as market trends or stock performance

(Géron, 2019, pp. 4-6; Kuhn & Johnson, 2018, p. 2), and even predicting user engagement on social platforms (Meta, 2025c; Narayanan, 2023, p. 30). ML performs particularly well in dynamic environments with non-stationary data, in problem areas where traditional algorithms fail, and in large-scale data analyses that reveal hidden knowledge (Géron, 2019, p. 7).

The core function of ML algorithms is to analyze data patterns while addressing uncertainties in the data. ML includes various methodologies that can be categorized through multiple taxonomies. Essentially, these algorithms need structured input data to solve specific problems, with the main classification based on their learning paradigm, which can be supervised, unsupervised, semi-supervised, or reinforcement approaches (Burkov, 2019, p. 3; Murphy, 2012, pp. 1-3). The next section will provide a clear overview of these paradigms.

2.9.1 Supervised learning

Supervised learning is a ML approach that uses both input data and corresponding labeled output data. A common example is spam detection, where the system learns to classify emails as either spam or legitimate based on previously labeled examples (Géron, 2019, p. 7). The classification process depends on historical data and current observations, using variables known as features. For email, such features might include the sender's address, the subject line, or the message content (Burkov, 2019, p. 3). For binary classification tasks like spam detection, logistic regression is often used because it can assign a probability to one of two possible outcomes (Murphy, 2012, pp. 21–23). Conversely, linear regression predicts continuous numeric outcomes (Géron, 2019, p. 9). The applications of supervised learning span different fields, including flower classification from images, handwriting recognition, and facial recognition systems (Murphy, 2012, pp. 6–9).

2.9.2 Unsupervised learning

Unsupervised learning is a branch of ML that works with data lacking labeled outcomes. Unlike supervised learning, where the model is trained on input-output pairs, unsupervised learning algorithms aim to find hidden structures, relationships, or patterns within datasets without prior knowledge of categories or labels (Géron, 2019, pp. 10-11; Murphy, 2012, p. 2). The main goal is to make sense of data by grouping similar instances or identifying unusual patterns. A common example of unsupervised learning is clustering, where algorithms like k-means or hierarchical clustering are used to divide data into meaningful groups based on similarity across multiple features. The aim is to find meaningful clusters and assign each data point to a specific cluster (Géron, 2019, pp. 10-12; Murphy, 2012, pp. 10-11). Another common task is dimensionality reduction, which simplifies complex data. The goal is to, for example, reduce features that are highly correlated and combine them into a new single feature that captures most of the information from the original features. This can be achieved using Principal Component Analysis (PCA) (Burkov, 2019, pp. 15-16;

Géron, 2019, pp. 11-12; Murphy, 2012, pp. 11-13). Finally, anomaly detection identifies data points that significantly differ from the rest of the dataset. The purpose is to remove these outliers before further processing (Burkov, 2019, p. 18; Géron, 2019, pp. 12-13).

2.9.3 Semi-supervised learning

This approach combines supervised and unsupervised learning by using a smaller set of labeled data and a much larger set of unlabeled data. With the help of the labeled data, this method aims to label all the unlabeled data. The algorithm detects patterns in the labeled data and applies the learned rules to the unlabeled data (Burkov, 2019, p. 4). A practical example could be classifying smartphone pictures uploaded to a cloud, which observe friends and family and group images where a specific person, like the uploader's mother, appears. At this point, the data is unlabeled. The system asks the user if the observed individual is the mother, and if confirmed, it labels the picture. The system then learns from this labeled data and can automatically label the remaining pictures (Géron, 2019, pp. 12-13).

2.9.4 Reinforcement Learning

Reinforcement learning is a unique approach within ML where an agent learns to make decisions by interacting with an environment and getting feedback in the form of rewards or penalties. The model aims to develop an independent understanding of the environment and simultaneously create a policy that can be compared to the input-output method used in supervised learning. By pursuing the highest reward, the model continually optimizes itself (Burkov, 2019, p. 4; Murphy, 2012, p. 2).

2.9.5 Machine learning in social bot detection

Previous studies have shown that ML techniques have great potential for identifying bots on social media platforms. Compared to traditional methods like crowdsourcing or rule-based heuristics, ML approaches handle key issues related to scalability, efficiency, and automation. As Habib et al. (2024, pp. 1–2) argue, ML has played a significant role in improving the detection of synthetic behavior online by allowing the systematic analysis of large interaction datasets.

Despite these methodological advances, several limitations still exist in the current research landscape. One key issue is the heavy focus on X, which limits the ability to apply findings to other platforms with different technical structures and user behaviors (Aljabri et al., 2023, p. 19). Additionally, most existing studies use supervised learning techniques. This trend is partly due to the easier access to X data, which enables researchers to manually label training sets or build on established datasets, such as those by Cresci et al. (2015, 2017; Aljabri et al., 2023, pp. 3–19). In comparison, research involving unsupervised ML is relatively limited

and mainly focuses on platforms like X, Facebook, and Instagram. Studies using semi-supervised learning are even less common, with current applications primarily on X and Weibo, revealing a significant research gap (Aljabri et al., 2023, pp. 3–23). Although some studies have explored synthetic interaction on Instagram, they mainly depend on supervised models and are confined to a small set of metric-based features (Akyon & Kalfaoglu, 2019, p. 2; Meshram, 2021, p. 121). Approaches that combine different feature types, such as profile metadata, text input, and visual content, have mostly been developed for X. For example, Goyal et al. (2023, pp. 6–7) propose a model that includes profile metrics, natural language analysis of tweets, and image-based data like profile pictures.

In the following chapter, the methodology of this research will be detailed to show how this study stands out from previous work. Several aspects will be discussed, including the framework that supports the thesis with a standardized outline and the specific scope of social media, from the platform itself to certain observed areas.

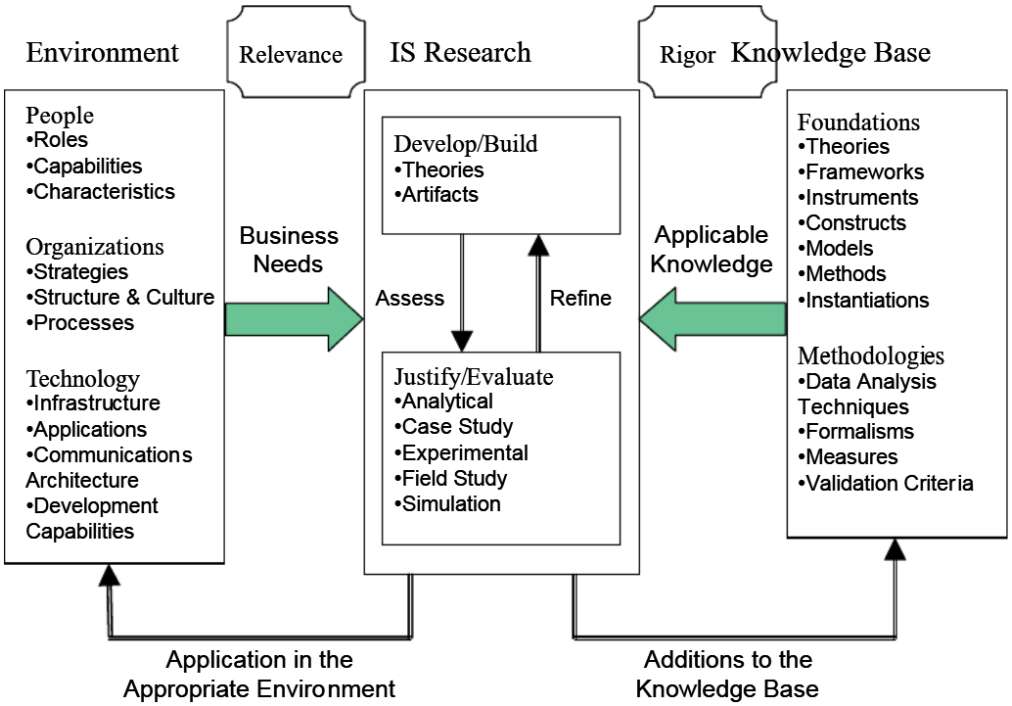
3 METHODOLOGY

This chapter outlines the methodological foundation of this thesis. It introduces the overall research framework, the processes applied to prepare and analyze the data, and the scope of the modeling approach.

3.1 Design Science Research

To establish a clear connection between the literature discussed in Chapter 2 and the practical component of this master's thesis, the DSR framework, as outlined by Hevner et al. (2004), is employed. This framework offers a scientific basis for building and assessing an artifact, which in this case is the predictive model for bot detection that serves as the main outcome of this thesis (Hevner et al., 2004, pp. 76-85).

Figure 6: Design Science Research Framework



Source: Hevner et al. (2004, p. 80).

The DSR process, shown in Figure 6, is based on existing knowledge. This knowledge base, created during the literature review, includes the fundamentals of social media, such as its characteristics, main providers, and underlying algorithms. The DIT, potential bot use cases, and current research on bot detection expand the theoretical foundation with additional focus on ML techniques that are central to the proposed artifact (Hevner et al., 2004, pp. 76-85).

At the same time, the Environment sets the context and importance for the research. It includes the business needs and economic drivers for effective bot detection, which are explained in Chapter 4.1. These needs define a problem space from which the research requirements are developed (Hevner et al., 2004, pp. 76-85).

The development of the artifact is an iterative process where the environment, the knowledge base, and the Information System (IS) research itself are interconnected. Throughout this process, the evolving artifact will be continually reviewed and assessed against the knowledge base to maintain scientific rigor and against the business needs outlined in the environment to ensure practical relevance. The specific evaluation methods for the artifact will be detailed later in Chapter 4.6 (Hevner et al., 2004, pp. 76-85).

3.2 Plattform

In this thesis, the focus is placed on the Instagram platform. One main reason for this choice is that, as mentioned, there is still relatively little academic research on bot detection specific to Instagram. Most existing studies focus on X, caused by the reasons addressed in Chapter 2.9.5. Instagram, on the other hand, works quite differently as it is a visually driven platform with its own user behavior patterns and interaction formats, which makes it an interesting and important area to explore.

As one of the largest social media platforms worldwide, Instagram plays a key role in shaping online communication and influence. It is also highly important from a business standpoint: after Facebook, Instagram provides the highest return on investment for marketing campaigns, which is why many companies and influencers rely on it heavily (Statista, 2024, p. 67). This popularity might make it a prime target for bots attempting to boost engagement or manipulate visibility. Studying bots on Instagram helps fill an important research gap and offers valuable insights to improve content authenticity and platform trust. Additionally, the challenges of working with Instagram data, such as limited access and fewer tools compared to other platforms, make it a subject worthy of closer academic examination.

Additionally, the feature design will not be limited to profile metric-based indicators, as seen in many previous studies. Instead, it will utilize Georges' identity layer model, which includes declarative, active, and calculated aspects of digital identity (Georges, 2011, pp. 39-41). This framework enables a more comprehensive assessment of user behavior by incorporating quantitative metrics, self-presentation, and social interactions.

3.3 Multimodal machine learning

Given the diversity of the extracted data, a multimodal learning framework is particularly well-suited for this thesis. Multimodal ML involves combining different types of data, known as modalities, into a single model, usually for classification or prediction (Burkov, 2019, pp. 100–102). Similar to human perception, modalities can be thought of as separate sensory channels: for example, words can be read or heard, while images are perceived visually (Baltrušaitis et al., 2018, p. 423). Many ML applications, including those for bot detection, still depend on unimodal input, meaning they utilize only one data type (Sleeman et al., 2022, pp. 1–2). In contrast, multimodal learning provides clear benefits. It can enhance robustness, especially when one modality is noisy or incomplete, and it enables different data sources to complement and strengthen each other (Baltrušaitis et al., 2018, p. 434).

In this thesis, three main data types will be integrated, based on the identity layer model introduced by Georges (2011). First, numerical features like follower count or post frequency belong to the calculated identity layer. Second, visual elements such as profile pictures or shared media represent the declarative identity layer. Third, textual information,

including user comments, pertains to the active identity layer. This layered and diverse input structure aligns with the principles of multimodal learning and offers a richer basis for classifying synthetic behavior on Instagram.

3.4 Semi-supervised machine learning

Given the multimodal nature of the dataset and the practical constraints in obtaining fully labeled training data, this thesis adopts a semi-supervised learning approach. In many real-world scenarios, and especially on platforms like Instagram, acquiring a large, reliably labeled dataset where bots and humans are classified accordingly is both time-consuming and difficult to scale. Unlike X, where labeled bot datasets are more readily available, Instagram provides significantly fewer pre-annotated resources. As a result, supervised learning methods that rely on abundant and verified labels become less practical in this context. Semi-supervised learning overcomes this limitation by using a small amount of labeled data along with a much larger pool of unlabeled data.

3.5 Tools

The practical implementation of this thesis was backed by a well-known data science stack built around the Python programming language. Python was chosen because of its strong ecosystem for data analysis, ML, and automation.

For data collection, especially from Instagram, the browser automation tool Selenium was used to simulate user interactions, allowing access to data beyond the limits of official APIs. The structured handling and processing of scraped data were managed using the library Pandas, while ML tasks such as feature transformation, clustering, and classification were carried out with scikit-learn. To analyze visual data, particularly for profile pictures, OpenCV was employed.

3.6 CRISP-DM

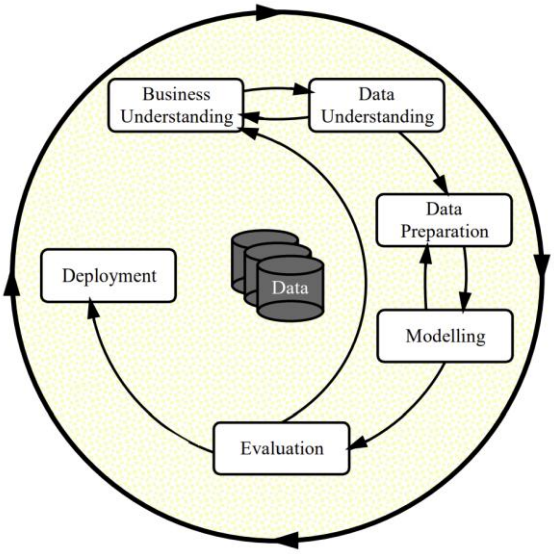
The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a widely used process model designed to guide data-driven projects through a structured and iterative framework. In this thesis, CRISP-DM is employed because of its industry-neutral and tool-free nature, making it highly flexible for both academic and practical use. The model helps organize the project's structure, ensuring that each phase of the data mining lifecycle is clearly outlined, traceable, and reproducible (Wirth & Hipp, 2000, pp. 1–2).

One of CRISP-DM's main strengths is its hierarchical structure. At the highest level, the methodology is divided into six stages, which represent the key phases of a data mining project. Each stage includes a set of general tasks that are widely applicable and can be adapted across different domains. These tasks are then further tailored and made specific

based on the project's unique needs. At the most detailed level, the methodology considers individual operations, decisions, and outputs, ensuring a clear and thorough workflow (Wirth & Hipp, 2000, p. 3).

The phases of the model are depicted in a closed circle, which also indicates the iterative nature of a data mining project, as shown in Figure 7. The project's output may not be the final phase, as there is room for ongoing improvements and the identification of future needs (Wirth & Hipp, 2000, pp. 4-5).

Figure 7: The phases of the CRISP-DM model



Source: Wirth & Hipp (2000, p. 5).

The phases of the model are explained in the following:

Phase 1: Business Understanding

The first phase of the CRISP-DM model emphasizes clarifying the overall project goals and translating them into a concrete data mining task, accompanied by a preliminary plan to guide the following analytical steps (Wirth & Hipp, 2000, p. 5).

Phase 2: Data Understanding

The data understanding phase involves collecting and exploring the available dataset to assess its structure, quality, and potential for uncovering meaningful patterns. It is closely tied to the initial problem definition, as forming a concrete analytical approach relies on early familiarity with the data (Wirth & Hipp, 2000, p. 5).

Phase 3: Data Preparation

The data preparation phase includes all steps needed to turn raw input into a structured dataset suitable for modeling. This involves tasks as cleaning the data or engineering new features. This is often an iterative and non-linear process (Wirth & Hipp, 2000, pp. 5-6).

Phase 4: Modelling

In the modeling phase, appropriate algorithms are chosen and applied, with parameter tuning carried out to improve performance. This phase remains strongly interconnected with data preparation, as the modeling process often uncovers underlying data limitations or the need for additional features (Wirth & Hipp, 2000, p. 6).

Phase 5: Evaluation

Once one or more models demonstrating promising analytical performance have been developed, this phase focuses on a comprehensive evaluation of their effectiveness. The goal is to ensure that the modeling process has adequately addressed the project's objectives and that no critical aspects have been overlooked (Wirth & Hipp, 2000, p. 6).

Phase 6: Deployment

Model development typically does not mark the conclusion of a data mining project, as the insights generated must be structured and delivered. Depending on the goal, deployment may involve anything from a basic report to a full data mining pipeline, typically run by end users (Wirth & Hipp, 2000, pp. 6-7).

A semi-supervised ML approach is proposed in order to enable the classification of larger quantities of unlabeled account data by leveraging a smaller manually labeled dataset. This shall reduce annotation efforts while maintaining performance (Aljabri et al., 2023, p. 23).

3.7 Ethical considerations

As this thesis involves the collection and analysis of user-generated data from Instagram, ethical considerations are of central importance. Only publicly available information was collected, and no attempts were made to access private accounts or bypass platform restrictions. The scraping process was designed to minimize server load and respect platform guidelines, and the collected data was anonymized where possible to prevent any identification of individual users.

A further ethical concern arises from the labeling of accounts as "bots" or "humans." Since false classification may affect how accounts are perceived in research or practice, results are presented in an aggregated and academic manner without singling out specific users. In addition, the limitations of the models are transparently acknowledged to avoid overestimating their accuracy or reliability.

Finally, this thesis recognizes that bot detection research can have dual-use implications: while it can improve platform safety, it may also inform the development of more sophisticated bots. To mitigate this, methodological details are provided in a way that supports academic reproducibility while avoiding unnecessary technical exposure that could be misused.

4 PRACTICAL IMPLEMENTATION

Chapter 4 describes the practical implementation of this research. Building on the theoretical foundation, the work follows the principles of DSR along with the CRISP-DM process model to organize the development and evaluation of the bot detection method. This chapter therefore connects the methodological framework with the practical steps taken to design, develop, and improve the solution.

4.1 Business understanding

The problem that motivates this work is the scarcity of prior research on Instagram, particularly in the areas of semi-supervised and unsupervised learning. The thesis aims to close this gap and address the issue of the underrepresentation of scientific work on the platform Instagram.

Instagram, as outlined in Chapter 3.2, plays a central role in modern online marketing. According to recent industry reports, it ranks second only to Facebook in terms of return on investment for social media marketing campaigns. While some automated accounts are used positively, for example, in automating replies to direct messages, many are deployed to artificially inflate metrics such as follower counts or engagement rates (Ferrara et al., 2016, p. 96).

This manipulation presents a major risk in influencer marketing, where choices are often made based on an influencer's perceived audience size and engagement levels. For example, if an influencer has 100.000 followers but a large part of them are bots, then companies' investments in collaborations may not produce the expected results.

Given the increasing reliance on influencer metrics for campaign planning, being able to verify the authenticity of engagement is crucial. Manually checking each follower or interaction isn't feasible on a large scale, so using ML algorithms for detection is an effective and scalable solution. This thesis aims to create a model that can classify Instagram engagement as either fake or real, contributing to both academic research and practical use.

Since the main goal of this thesis is to detect synthetic engagement on Instagram using ML, the next step is to gain a detailed understanding of the available data. To effectively design

and implement an appropriate model, it is important to analyze the structure, quality, and characteristics of the data that will guide the analysis. This includes identifying which features are available, how reliable they are, and whether they provide enough insight into user behavior to help detect bots. The following section describes the data sources, data collection methods, types of data gathered, and initial observations relevant to the modeling process.

4.2 Data Understanding

The following section deals with the data understanding phase. At this stage, the collected data is explored and described in order to get a first impression of its structure and content.

4.2.1 Data collection

The data in this thesis includes multiple modalities as described in Chapter 3.3. These modalities consist of numerical features like profile metrics (e.g., number of followers, number of followings, number of posts), combined with visual features such as profile pictures. The data was initially collected by scraping comments from selected Instagram posts. From these comments, each username was extracted and then used to gather additional information about the profiles.

The selected domains for observation cover three areas on Instagram where bot activity is likely to be significant, based on the knowledge base established in Chapter 2.7.2 and 2.7.3: meme pages, financial information pages, and accounts of politicians. Meme pages were included because their high engagement levels make them appealing for bots to artificially inflate interactions. Financial pages were chosen because bots are often used in this context to promote investment schemes or dubious financial advice. Political accounts were selected because previous studies have shown that bots can influence opinion formation and increase polarization. In each domain, comment sections from multiple pages were scraped to gather a diverse dataset.

Appendix 4 provides an overview of the accounts from which comments were collected. The distribution across the three domains is uneven, as meme pages generally generate higher engagement and more comments per post. In contrast, political and financial accounts required a broader selection of pages to reach a similar number of entries. When choosing the pages, priority was given to accounts that appeared credible and to avoid those that seemed suspicious or malicious. Posts selected for data extraction were based on engagement levels, with at least 200 comments and no more than 1.500 per post to ensure diversity and prevent overrepresentation of single posts. Giveaways or similar posts were excluded because they tend to attract repetitive, homogeneous comments. Replies to comments (nested comments) were also excluded to maintain consistency, with only initial comments considered.

The scraping process was carried out using Python. The Selenium library was used to automate the Chrome browser, log in to Instagram, and navigate to the desired pages. Selenium finds website elements based on their ID, CSS selectors, or XPath. For each selected post, the script detected elements containing the needed information: the username (h2-element), the comment text (h3-element), and the timestamp (time-element). Figure 8 shows an example of the scraping script.

In total, approximately 13.000 comments were initially collected. After removing unusable entries (such as GIFs, replies, or incomplete data), the dataset contained approximately 10.000 valid entries. Further cleaning involved the removal of duplicate usernames and entries where values could not be extracted due to occasional loading failures, even after implementing sleep timers. Fortunately, this represented less than 0.5% of the dataset.

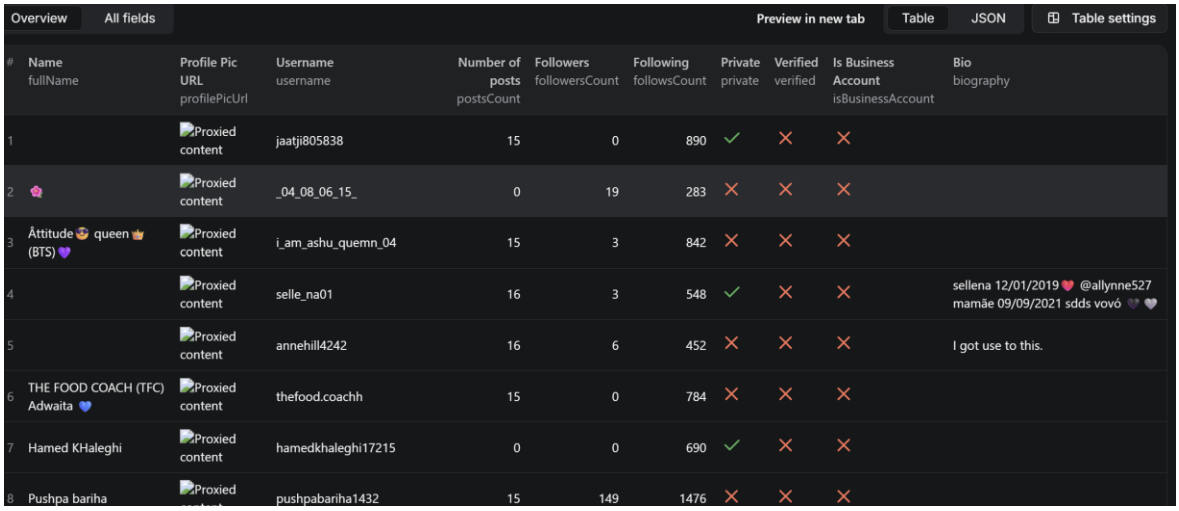
Figure 8: Python script to extract comments

```
for container in container_elements:
    # Searching the comment-element in the container and extracting the text
    try:
        text_div = container.find_element(By.CSS_SELECTOR, "div.xt0psk2 span").text
    except:
        text_div = "[No xt0psk2 found]"

    # Searching the username-element in the container and extracting the text
    try:
        heading = container.find_element(By.TAG_NAME, "h3").text
    except:
        heading = "[No h3 found]"

    # Searching for the time-element and extracting the datetime
    try:
        time_element = container.find_element(By.CSS_SELECTOR, "time._a9ze._a9zf")
        timestamp = time_element.get_attribute("datetime")
    except:
        timestamp = "[No timestamp found]"

    # Matching the extracted information to the attributes which will be written into the csv-file
    results.append({
        "heading": heading,
        "timestamp": timestamp,
        "text_div": text_div.strip().replace("\n", " ").replace("\r", " ")
    })
```

Source: Own work.

The cleaned usernames were then categorized into three domains (“Memes,” “Finance,” and “Politics”) and used to gather profile data. This was done using the service Apify.com and its “Instagram Profile Scraper” tool, which allows bulk input of usernames and outputs multiple profile features. Figure 9 shows an example of the tabular output. To process the data locally, JSON files were downloaded and then converted into a CSV files using Python scripts. During this step, formatting issues (such as undesired line breaks caused by “/n” values) were fixed, ensuring that only 17 of the 24 available dimensions remained for further analysis.

Figure 9: Apify output of the profile information



Overview | All fields | Preview in new tab | Table | JSON | Table settings

#	Name fullName	Profile Pic URL profilePicUrl	Username username	Number of posts postsCount	Followers followersCount	Following followsCount	Private private	Verified verified	Is Business Account isBusinessAccount	Bio biography
1		Proxied content	jaatji805838	15	0	890	✓	✗	✗	
2		Proxied content	_04_08_06_15_	0	19	283	✗	✗	✗	
3	Âttitude queen (BTS)	Proxied content	i_am_ashu_quemn_04	15	3	842	✗	✗	✗	
4		Proxied content	selle_na01	16	3	548	✓	✗	✗	sellena 12/01/2019 @allynne527 mamãe 09/09/2021 sdds vovó
5		Proxied content	annehill4242	16	6	452	✗	✗	✗	I got use to this.
6	THE FOOD COACH (TFC) Adwaita	Proxied content	thefood.coachh	15	0	784	✗	✗	✗	
7	Hamed KHaleghi	Proxied content	hamedkhaleghi17215	0	0	690	✓	✗	✗	
8	Pushpa bariha	Proxied	pushpabariha1432	15	149	1476	✗	✗	✗	

Source: Own work.

Finally, profile pictures were retrieved using a Selenium-based script. The script navigated to each user's profile, found the element with the profile picture, and saved it to a local repository using the username as the filename. This ensured a clear link between images and user accounts.

4.2.2 Gathering of labelled data

As discussed in earlier chapters, one of the main challenges in bot detection research on Instagram is the lack of labeled datasets. To address this, several strategies were implemented during this thesis. Researchers from previous studies were contacted for access to labeled datasets, but no responses were received within the given timeframe. Public datasets available on GitHub and Kaggle were downloaded, but these datasets typically focused on only one modality and did not match the multimodal approach of this thesis.

To independently generate labels, existing features from the Apify service were reviewed. One useful attribute was the "verified" badge, which signals authentic accounts. According to Meta, accounts can earn a verification badge if they represent a public figure, brand, or organization meeting criteria such as notability, authenticity, and identity verification (Meta, n.d.; Meta, n.d.). Of the 11.500 entries, captured by Apify, 267 were verified and thus could be confidently identified as genuine accounts. Additionally, a small survey, observable in Appendix 2 , was conducted among the author's Instagram followers, resulting in 50 more genuine accounts. In total, 317 accounts could be reliably labeled as genuine.

For bot accounts, a different approach was necessary. Since no datasets with labeled bots were publicly available, a controlled experiment was conducted. Using the service *buzzoid.com*, fake followers were purchased for a private Instagram account controlled by the author. In Appendix 3, the services provided by the website are visible, with the first

option, ‘High Quality Followers’, chosen to ensure that the purchased accounts are not ‘real’, as other services on the platform differentiate themselves through this feature. These followers, which serve no purpose other than inflating metrics, could then be collected and labeled as bot accounts. While some of these accounts mimicked human users with profile pictures and posts, closer inspection revealed clear patterns of inauthenticity. Many users exhibited engagement issues on their posts, used usernames with suspicious numeric sequences, or followed hundreds of accounts while having very few followers themselves. These characteristics are also consistent with the indicators identified in previous studies, such as Cresci et al. (2015, pp. 59–63).

All usernames of genuine and bot accounts were processed through the same pipelines described earlier (Apify scraping, local cleaning, and profile picture collection), resulting in a labeled dataset that could be used for training and evaluation.

4.3 Data understanding

The whole data collection process followed a structured pipeline. First, comments were scraped and cleaned; then, profile information was gathered, and finally, profile pictures were added. At each stage, incomplete or faulty entries were removed to ensure consistency. In the end, all three modalities were aligned through the username as a unique identifier.

As mentioned before, around 13.000 comments could be scraped and saved locally. Out of this dataset, approximately 11.500 entries could be captured by the Apify platform. The transformed CSV file, which includes all remaining scraped profiles, inherited entries containing crucial information, such as followerCount and followingCount, that could not be captured. These entries were also removed because those were essential features. Fortunately, this affected only around 20 entries in the total dataset. The usernames of the remaining 11.500 entries were used to execute the profile scraping script. This procedure was the most seamless, resulting in a loss of around 100 profiles from accounts that were deleted in the meantime.

This process reduced the dataset to slightly over 11.000 entries, of which 317 were labeled as genuine and the same amount labeled as bots. The remainder of the dataset remained unlabeled to be be used in the semi-supervised learning framework.

Featues

After acquiring the data and all structural affordances, the dataset was ready to be observed. Each modality brings a different perspective on the proclaimed identity of an online user. The data extracted from the comment section inherits the attributes ‘username’, ‘comment’, and ‘timestamp’. The profile picture on the visual perspective of modalities is one attribute on its own. The attributes extracted from the profile itself have a much higher variety. In Table 1, the features and their description are listed accordingly.

Table 1: Extracted features

Comments	
Username	Id for the profile that wrote the comment (String)
Comment	The comment under one of the observed profiles (String)
Timestamp	The timestamp of publishing the comment (Datetime)
Visual	
Username	The username of the person with a particular profile picture (String)
Profile picture	The picture a user utilizes online (.png)
Profile	
Username	Username of the observed profile (String)
Verified	Stating if a user is officially verified by Instagram (True/False)
Private	Stating if the user has a public profile where everyone can see the content (True/False)
postCount	The number of media published by an account (Numeric)
joinedRecently	States if the account was created lately (True/False)
isBusinessAccount	States if the account has professional functions and advertises products (True/False)
igTVCount	The number of igTV videos published by an account (Numeric)
highlightedReelCount	The number of Reels (Short videos) published by an account (Numeric)
hasChannel	Stating if an account has a broadcast channel where they communicate information with certain followers (True/False)
followsCount	The number of accounts the profile is following (Numeric)
followerCount	The number of accounts that follow the profile (Numeric)
externalURL	Stating if there is an external URL included in the profile's description (True/False)
Bio	The text of the profile biography, which a user curates and publishes (String)

Source: Own work.

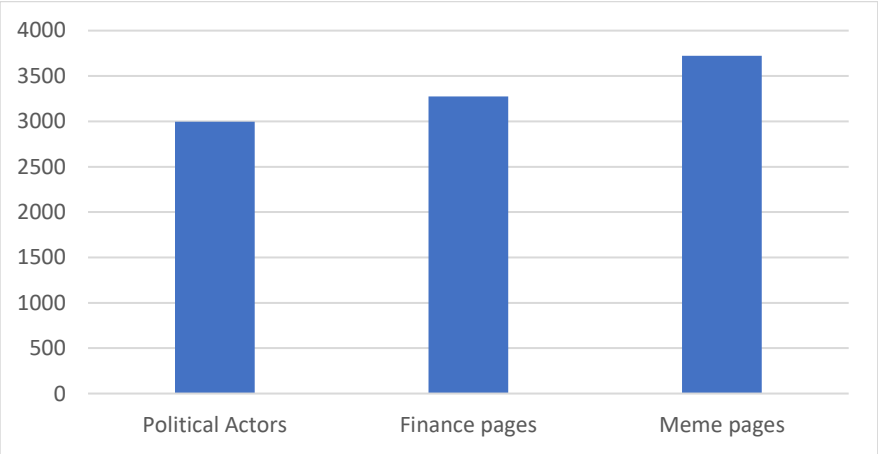
When excluding from the calculation, 15 attributes were extracted. The data can be categorized mainly into string, numeric, and boolean data, complemented by one datetime attribute and a PNG picture. If a user had the value ‘True’ in the verified attribute, the whole row was transferred to the training data. For the unlabeled dataset, the attribute was deleted because all remaining accounts had ‘False’ and no value could be generated from this.

Special emphasis was placed on the utilization of features in previous research to understand which feature modalities and individual features could benefit the model. Since many studies involving metric-based features achieved remarkable results, those features were given special attention to ensure that the artifact in this thesis also provides value. That is why those features are more represented than features from other modalities.

Descriptive statistics

After removing all profiles from the dataset that contained missing features, the dataset with unlabeled data consisted of 9.996 entries, and the dataset containing labeled bots and genuine users had 634 entries. This results in a challenging share of labeled data of roughly 6.3 percent in relation to the unlabeled data. The distribution of accounts coming from different domains can be seen in Figure 10.

Figure 10: Distribution of accounts by domain

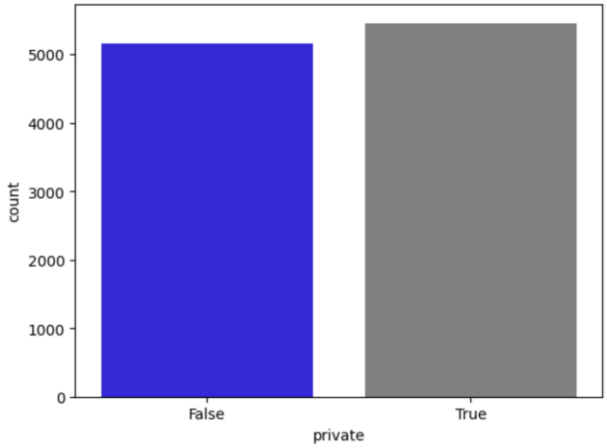


Source: Own work.

To get familiar with the data, the training data and a set of unlabeled entries were loaded into a Python script and analyzed using the Pandas library. Many of the attributes are not initially informative, such as the username or the biography the user published.

Initially, it was interesting to observe how the data is categorized into private and public domains. For the analysis and acquisition of data, there was no obstacle; all the data needed can be extracted if a user is either private or public. Figure 11 shows that both groups are represented nearly equally.

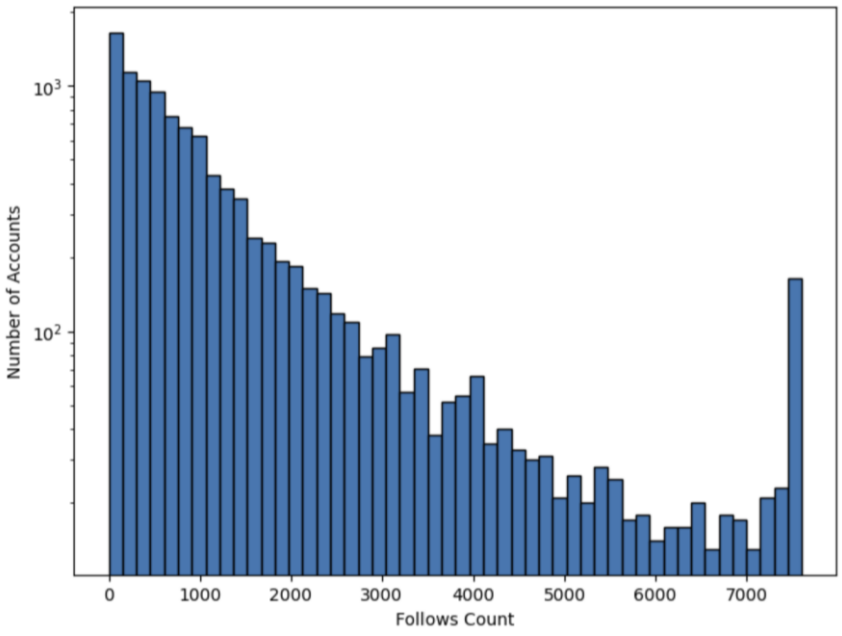
Figure 11: Comparison of private vs. public accounts



Source: Own work.

When observing the number of followers one account has, it is notable that the data is skewed to the right, with the mean at 1243 followers, compared to the scale that ends at 7610. Nevertheless, a significant portion of accounts, exceeding 7000, can be seen in the histogram displayed in Figure 12.

Figure 12: Distribution of Followings

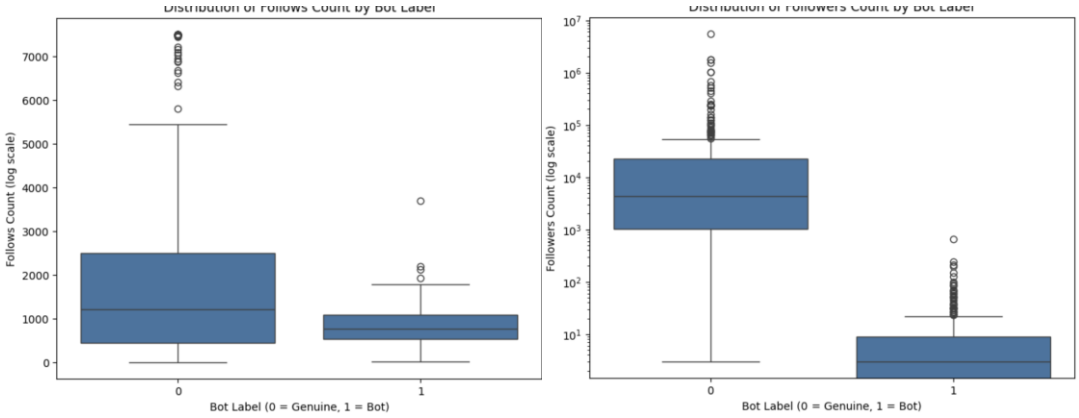


Source: Own work.

As shown in the left diagram of Figure 13, the rise in follows, which begins at 7.000, originates from human behavior. Bots on the other side rarely follow more than 1.082 accounts, and they also struggle to display a large number of followers on their accounts.

they are significant indicators of whether an account is a bot or not.

Figure 13: Comparison of follows and followers - labeled data

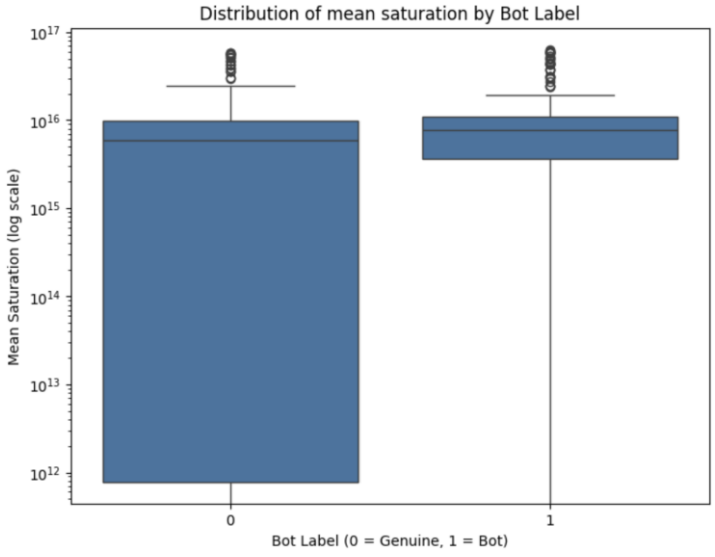


Source: Own work.

50% or fewer bots get only three followers, while genuine users get 4.377 in comparison. This might also be influenced by outliers who have millions of followers, which influences this statistic. When it comes to posts published, the quantiles of 25, 50, and 75 percent were at exactly 15 posts for the bots, while the 25 % to 75% quantile distance for genuine users was 899.

While observing the visual data, only a few things polarized. The feature where the saturation of the picture was calculated differed across both classes, as can be seen in the boxplots of Figure 14.

Figure 14: Mean saturation comparison



Source: Own work.

The promising feature ‘face_detected’, which indicated whether a face was visible on the profile picture, unfortunately, didn’t add much value. 205 faces were detected for the bot accounts, while for the genuine users, 215 were detected out of the 317 entries in each domain.

Since the comments were mostly scraped from unlabeled accounts, further categorization and data exploration cannot be performed here. The comment features are more important in later parts of this work, as the semi-supervised model will also learn from these features. The comment features will not be used in supervised training on labeled data because there are no clear labels for the accounts that commented.

4.4 Data preparation

In this chapter, all the measures are listed for manipulating the datasets and making them more suitable for the modeling task following this chapter.

4.4.1 Dealing with isnull

All datasets, including training and unlabeled datasets for visualization and profile metrics, were checked for cells without values. This was done via Excel to ensure every entry is complete and does not affect subsequent steps. The ‘username’ column was checked for duplicates to prevent any entry from influencing the model with greater weight. No duplicates or missing values were found in any of the datasets.

4.4.2 Dealing with irrelevant data

While exploring the attributes, it became apparent that the attributes ‘hasChannel’ and ‘joinedRecently’ both only hold the ‘False’ value, which means the model cannot learn from these columns since all entries in the training data are the same for these attributes. Because they do not provide any value, the attributes were removed from the data frame and not considered further for the model.

4.4.3 Feature engineering

To complement the data with insightful features that could not be scraped upfront, features were engineered.

Starting with the comments, the length of each comment was determined by measuring its character count, while the number of words was calculated by counting the spaces and adding one. The number of mentions and hashtags included in the comment was extracted by searching for ‘@’—the first character that appears when someone mentions another person—and similarly for hashtags by searching for ‘#’. Additionally, the comment was checked for ‘https:’ to identify if a link was pasted into the comment section. Lastly, it was determined whether the comment contained emojis. The extracted timestamp alone does not itself. The difference between these two timestamps indicates the user's response time to the post. A snippet of this data can be observed in Appendix 5.

For the visual data, features were also engineered. These were extracted using a Python script and the libraries cv2 and torchvision. With cv2, the picture was analyzed for its saturation, hue, and mean color value as observable in Appendix 6. The sharpness was measured, along with a feature indicating whether a face was detected in the picture. torchvision narrowed the initial image down to 2048 ResNet-Embeddings.

The profile metrics were also enhanced by new features. The username alone cannot contribute significantly to the model. Therefore, features were developed based on the length of the username and the number of characters it contains. This is very similar to the biography, as the text also doesn't contribute much. A binary feature was created that indicates whether a biography is present, since it is optional on the platform. The length of the biography was also measured.

Sentiment analysis of the biography and the user comments was not performed because the diversity of languages in the dataset would make this process significantly more complex.

The strings indicating whether a specific attribute is either true or false don't add much value to the model in the end. That's why these were converted into a numeric format where 'True' was replaced by '1' and 'False' by '0'. The affected attributes were 'private', 'isBusinessAccount', and 'externalUrl'. The attributes 'joinedRecently' and 'hasChannel' were not affected because they were dropped from the dataframe earlier.

To prepare the dataset for modeling, several steps were taken. First, the original dataset included a label indicating whether an account was a bot or genuine. Since this label should not be part of the features used for testing, it was separated and stored as the target variable.

Afterward, the data was standardized. This step was crucial because the values of different profile metrics vary greatly in scale; for example, the follower count can reach thousands, while other features like boolean attributes are only 0 or 1. Standardization made sure all features were transformed into a comparable range, which is especially important for models that are sensitive to feature scaling like Support Vector Machines and Logistic Regression.

Finally, the dataset was split into training and test sets. This allowed the models to be trained on one part of the data and then evaluated on unseen data. By keeping the test data separated during training, the evaluation results provide a more realistic view of how well the models are likely to perform on new accounts.

4.5 Modelling

Since all data preparation steps were completed, it was possible to build the model itself. For this thesis, several classifier models were considered to identify good baseline models for training the subsequent models. Because the goal of this work is to classify between social bots and genuine users, a few classifiers were taken into consideration. Starting with the

Random Forest Classifier, along with Support Vector Machines, Logistic Regression, and LightGBM, these models were evaluated based on their performance in classifying the labeled dataset, where bots were labeled with '1' and humans with '0'. The share of training and test data was 80 to 20, resulting in 507 entries for training the dataset and 127 entries for the test dataset. This split was consistent across all trained models. The dataset itself was perfectly balanced, containing exactly 50 percent bots and 50 percent humans.

Initially, each modality was trained separately, resulting in 8 different models, with each classifier trained on the metric data and the visual data. Each model employs a different approach to classifying the data. For example, the Random Forest establishes multiple subsets of the original dataset and builds decision trees, searching for suitable and strong features within each subset. This classifier reduces variance through diverse samples and also handles imbalanced data effectively (Burkov, 2019, pp. 9-10; Géron, 2019, pp. 199-201). The Support Vector Machine classifier is used to distinguish between, for example, two classes, using decision boundaries that can be visualized as lines separating the groups. It creates a multi-dimensional space where each entry finds its unique position, then observes this space to find the instance that separates the classes; the decision boundary (Burkov, 2019, pp. 5-8; Géron, 2019, pp. 155-163). Logistic Regression takes several numeric inputs in the form of features to estimate the probability that an entity belongs to a specific label, which is either 0 or 1. The result is a number between these two values (Burkov, 2019, pp. 6-8; Géron, 2019, pp. 144-146). Lastly, LightGBM is a classifier used for the deployment on large dataset for the purpose of classification or regression (GeeksforGeeks, 2025).

After running the models, initial results were observed and compared. However, before the final comparison of the specific classifiers, the models were tuned by adjusting the hyperparameters used in their setup. Each model allows certain parameters to be configured (Bentéjac et al., 2020, pp. 1938-1943). The adjustable parameters for each model are listed in Tables 2 to 5 below.

Table 2: Hyperparameter Random Forest

n-estimators	Number of trees
max_depth	The maximum depth the tree develops
min_samples_split	The number of minimum samples built out of the original data
min_samples_leaf	Minimum of samples per leaf

Table 3: Hyperparameter Support Vector Machines

kernel	Declares how the boundaries will be drawn
C	Regularization parameter
gamma	Kernel coefficient
probability	Enables probability estimates

Table 4: Hyperparameter Logistic Regression

max_iter	Maximum of iterations
C	Regularization parameter
penalty	Kernel coefficient

Table 5: Hyperparameter LightGBM

n_estimator	Declares how the boundaries will be drawn
max_depth	Regularization parameter
learning_rate	Kernel coefficient

Continuously adjusting the hyperparameters led to several results, which were evaluated using the performance measures introduced later in this chapter. The goal was to build each baseline model with parameters that best fit the model according to those performance metrics. The corresponding results of the baseline models show interesting insights, which can be seen in Table 6.

Table 6: Baseline-model results

Model	Accuracy	Precision	Recall	F1-score
Random Forest Metrics	0.97	0.97	0.97	0.97
LightGBM Metrics	0.95	0.95	0.95	0.95
Logistic Regression Metrics	0.95	0.95	0.95	0.95
Support Vector Machines Metrics	0.93	0.93	0.93	0.93
LightGBM Visuals	0.69	0.69	0.69	0.68
Support Vector Machines Visuals	0.66	0.73	0.66	0.64
Logistic Regression Visuals	0.65	0.63	0.70	0.66
Random Forest Visuals	0.62	0.67	0.62	0.59

Source: Own work.

As can be seen, the models trained on profile metrics such as the number of followers or followings show excellent results, with the Random Forest classifier correctly labeling almost every entry except two false negative and two false positive predictions. Additionally, LightGBM and Logistic Regression share second place, followed by the Support Vector Machine classifier. Regarding feature importance, the classifiers differ in how they weight certain features. The feature bioLength appears to be very significant, as all classifiers include it among their top three most important features. The postsCount feature also seems notably important, as it appears in the upper spectrum of importance for each classifier. On the other hand, the features followersCount and followsCount are crucial for the LightGBM and Random Forest classifiers but are less prominent in the other models. The highlightedReelCount feature is the most important indicator for the Logistic Regression model, while it ranks eighth in the LightGBM model.

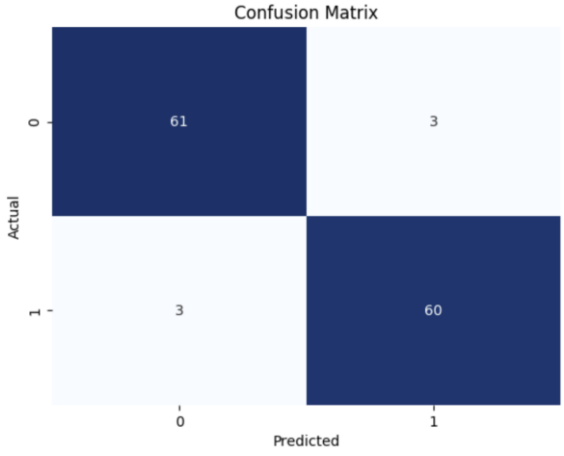
The results for the visual features are not impressive because they are quite low. The best model in terms of accuracy was the LightGBM classifier, which correctly labeled 87 entries, with 15 predicted as humans but actually being bots, and 25 predicted as bots but actually being humans.

By training all the baseline models individually, the goal was to achieve the best possible outputs since the classifier used in the semi-supervised model later on should be as robust as possible. A direct comparison with the knowledge base in Chapter 2.8 confirmed the development process, even though the initial accuracy results seemed questionable. Additionally, the relatively poor performance of the visual models was an obstacle. By tuning the models with hyperparameters, both the profile-metric-based and visual-based models could be improved iteratively. This showed how strongly individual classifiers perform on specific data modalities and provided guidance on how the modalities should be best fused accordingly to ensure the business needs are met.

4.6 Evaluation

After all the models were run, special focus was placed on evaluating their performance. Several measures can be used to determine whether the model performed well or not. First, the confusion matrix was utilized to get an overview of which cases were classified into which labels compared to the actual labels that the model was unaware of for the test data. Figure 15 shows such a matrix, which includes the fundamental numbers on which the following performance measures are built. In total, the model correctly labeled 101 entries.

Figure 15: Confusion matrix LightGBM metrics



Source: Own work.

The remaining six incorrectly labeled entries are evenly distributed between the False-Positive and False-Negative categories. Entries predicted as bots but are actually human users are labeled as False-Positives (FP), while those mistakenly identified as humans but

are actually bots are labeled as False-Negatives (FN). The other two categories represent True-Negative (TN) and True-Positive (TP), respectively.

One of the performance indicators is the precision of the model. It is calculated by dividing TP by the total number of positive labels the model predicted, which is obtained by adding TP and NP. The formula is as follows:

$$\frac{TP}{TP+FP} \tag{1}$$

The recall states how many labels were truly positive compared to the sum of true positives and the negative false ones, which is the total of positive entries in the test data.

$$\frac{TP}{TP+FN} \tag{2}$$

The accuracy is the sum of all TP and TN classifications compared to all the entries that were in the test data.

$$\frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

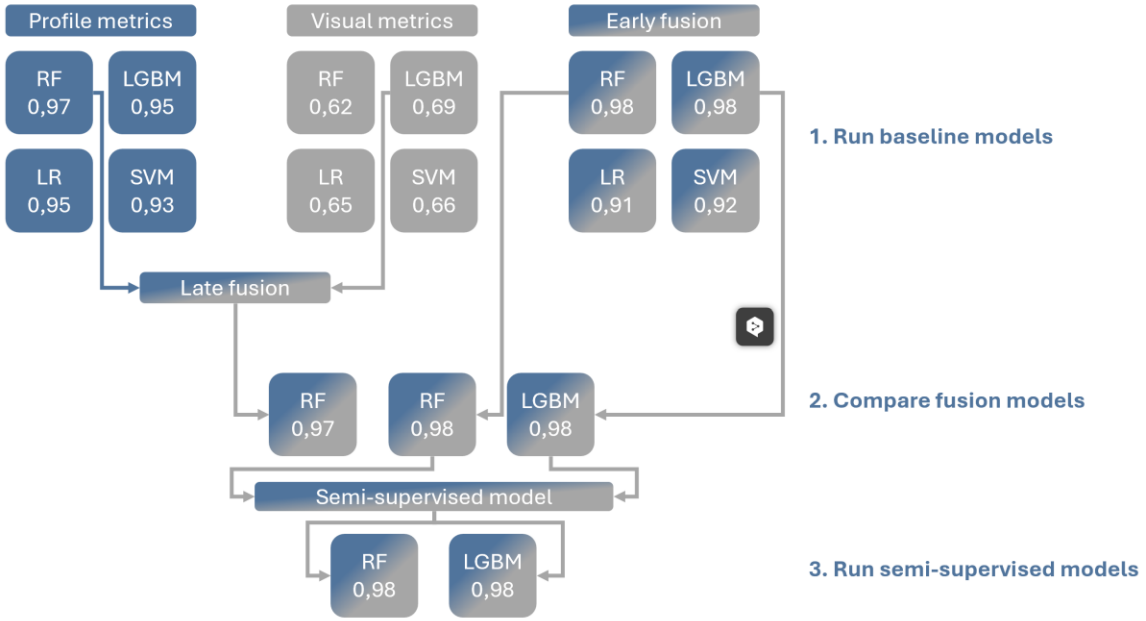
In other use cases, there might be a clear focus on one of the performance metrics. For example, in a system classifying spam emails, it would be advantageous to have a high precision score, as non-spam emails should be classified correctly, while it is acceptable to have a few spam emails in the inbox but not the other way around (Burkov, 2019, pp. 13-18).

Since this research aims to be as accurate as possible without favoring any specific class, the main focus is on achieving the highest overall accuracy of the model.

After successfully establishing the baseline models, the next step involves combining the insights from both models. This can be achieved either by concatenating all features into a single labeled dataset and training this combined data on the specific classifiers, or by integrating them after the initial baseline models, which differ in modality scope, have been trained.

Initially, the two best models from each domain were combined after running them separately. This combined result was then tested and reclassified using an additional meta-classifier, which in this case was the random forest classifier, since the profile metrics features appeared most important. The results were promising, but as expected, the visual features contributed little to the overall model. Nevertheless, an accuracy of 0.97 was achieved.

Figure 16: Final model evaluation



Source: Own work.

By combining both feature sets into a single dataset, early fusion models could be developed. They produced promising results, with the random forest classifier reaching an accuracy of 0.98, and the other performance metrics also achieving the same percentage. The second-best model was the LightGBM classifier as it can be seen in Figure 16, with only a one-percentage-point difference in Recall. This laid the groundwork for the next model, which should classify the unlabeled dataset in a semi-supervised manner.

4.7 Semi-supervised model

In the subsequent iterations, semi-supervised models were applied to classify the large volume of previously unlabeled data. The following subchapters present and discuss the results obtained from these models.

4.7.1 Iteration 1

The semi-supervised models were built with exactly those classifiers mentioned above. The labeled dataset served as the training data, while the unlabeled data was the test data, which was labeled accordingly with a confidence threshold of 0,95. Over 75 iterations, the model labeled the entries.

The results were promising but varied significantly between classifiers. While the random forest classifier was able to label 87.60% of the entries, the LightGBM model labeled

99.61% of all entries. Additionally, the distribution of bots and human labels differed greatly when comparing the outputs of both models, as shown in Table 7.

Table 7: Performance evaluation of semi-supervised models Version 1

	Human	Bot
Random Forest	8576 (98,0 %)	179 (2,0 %)
LightGBM	8181 (82,2 %)	1774 (17,8 %)

Source: Own work

At first, this seemed crucial because the models had identified different patterns to tell the two classes apart.

Random Forest tends to be more stable, while LightGBM makes bolder assumptions about the data. This difference was also evident in the confidence levels of both classifiers. LightGBM only classified 38 entries with a confidence below 0,95, whereas Random Forest was less confident than 0,95 in 1239 cases. The importance of features also varied, with the top 10 features for the models being as follows:

Figure 17: Feature importance Random Forest and LightGBM, Iteration 1

Feature	Importance
followersCount	0.189123
postsCount	0.120869
bio length	0.117661
mean_value	0.079931
numbers in username	0.079475
mean_saturation	0.052780
mean_hue	0.051167
highlightReelCount	0.044747
followsCount	0.036505
igtvVideoCount	0.035852
sharpness	0.026035
externalUrl	0.019377
private	0.015293
length username	0.014358
face_detected	0.005960
isBusinessAccount	0.005682

Feature	Importance
followersCount	20.89
followsCount	11.22
sharpness	10.99
bio length	10.17
postsCount	8.62
mean_value	8.51
numbers in username	8.51
mean_hue	6.36
length username	5.27
mean_saturation	4.32
private	3.99
face_detected	0.37
igtvVideoCount	0.36
highlightReelCount	0.29
externalUrl	0.11
isBusinessAccount	0.02

Source: Own work.

Figure 17 shows the feature importance of the Random Forest model on the left and the LightGBM feature importance on the right. It is visible that visual features also add value to the classification even when the balance cannot be fully maintained. In the model with Random Forest as the classifier, they contributed around 21%, while in the LightGBM model, about 30% is contributed through visual features.

The next step was to determine how many entries were classified the same way by both models, which totaled 8.680 entries. Then, this was narrowed down to a subset where the labels matched and both models had a confidence score of 0,99 or higher. This reduced the dataset to 7.619 entries, which were complemented by the existing 634 entries of the labeled data used before. Since the baseline training data did not have any features related to

comments, imputation was used to synthesize these values based on the now freshly labeled dataset after the first run of the self-learning model. This resulted in a new training dataset with a total of 8.253 entries and a new test dataset of the remaining entries where both models could not agree on a label with a confidence of 99 %, totaling 2.374 entries.

As a further improvement, the comments now complement the entire classification process. Since the labeled dataset now includes 8.253 entries and the test data 2.374, there is a proper ratio between known and unknown labels of 71,3% and 28,7%. To assess how the comments contribute to the process, both semi-supervised classifier models were run again.

4.7.2 Iteration 2

The models could incorporate the comment features since they contributed at least 16 % to each model, as can be observed in Figure 18, where the left-sided output is from the RF classifier and the right-sided from the LightGBM classifier. When comparing the labels of both classifiers, it was notable that the agreement on a label with 99 % confidence decreased significantly, from 76% in the initial run to 48% in the second run with the new labeled data. A small number of data entries, totaling 1.233, remain unlabeled.

Figure 18: Feature importance Random Forest and LightGBM, Iteration 2

Feature	Importance
followersCount	18.05
bio length	10.53
mean_value	9.13
postsCount	8.38
Response time in minutes	8.20
followsCount	7.47
numbers in username	6.22
mean_hue	6.18
sharpness	4.96
Length character	4.52
mean_saturation	4.13
length username	3.27
# words	3.10
private	3.04
Emojis?	0.88
face_detected	0.73
igtvVideoCount	0.46
highlightReelCount	0.41
externalUrl	0.30
isBusinessAccount	0.03
# mentions	0.01

Feature	Importance
followersCount	0.164403
bio length	0.087062
postsCount	0.084890
Response time in minutes	0.083119
mean_value	0.064596
numbers in username	0.056245
Length character	0.053403
mean_hue	0.049931
# words	0.038939
mean_saturation	0.036951
followsCount	0.033893
highlightReelCount	0.033661
igtvVideoCount	0.032569
Emojis?	0.021525
sharpness	0.020210
private	0.015007
externalUrl	0.012395
length username	0.011527
face_detected	0.007157
# mentions	0.003493
isBusinessAccount	0.003350

Source: Own work.

4.7.3 Iteration 3

In the last iteration, the newly labeled entries were added to the previously labeled data set of 8.253, increasing it to 9.394 entries. The models no longer operate independently but are combined into a cooperative learning model where decisions are made using signals from both classifiers.

The results show that the differences in labeling the remaining entries increase. By combining the confidence scores of both classifiers and calculating the mean, only 13% of the labeled data after the run is labeled with a confidence score of 99% or higher. While the LightGBM classifier tends to label an entry as a bot more easily, the Random Forest classifier is only confident about one additional entry being a bot.

To preserve the quality of the labeled data, the classification stops after this iteration, and the overall results are observed.

1.066 entries remain unlabeled confidentially. 8.928 have been successfully labeled with a confidence level of 99 % or higher. Out of this labeled dataset, 109 entries were confidently identified as bots, which accounts for 1,28 % of all confidently labeled data. The remaining 98,72 % were labeled as humans.

In contrast, including all labels with a confidence level of 80% or higher would add 490 extra bot-labels, resulting in a total of nearly 6% of all labeled data. The effects of different confidence levels are shown in Table 8.

Table 8: Confidence Comparison

Confidence Level	Labeled Data	Share Human	Share Bots
99 %	8.928	8.820 (98,72 %)	109 (1,28 %)
95 %	9.445	9.234 (97,77 %)	211 (2,23 %)
90 %	9.880	9.377 (94,91 %)	503 (5,09 %)
85 %	9.976	9.384 (94,07 %)	592 (5,93 %)
80 %	9.982	9.384 (94,01 %)	598 (5,99 %)

Source: Own work.

When comparing the model's ability, including labels with a confidence of 80% or above, almost all entries could be labeled as such, with only 14 remaining in the 50% to 79% confidence range. Thus, the semi-supervised model shows that such method maintains label quality while scaling up with the dataset.

5 DISCUSSION

This chapter reflects on the main outcomes of the research and places them in a broader academic and practical context. It discusses the key findings in relation to the research questions and prior studies, draws overall conclusions from the developed models and methodological approach, and outlines promising directions for future research.

5.1 Findings

The modeling results of this thesis demonstrate a clear trend: classifiers trained on profile metrics consistently outperform those relying solely on visual features. This aligns with prior studies, which have emphasized the importance of structural and behavioral indicators such as follower/following ratios, posting frequency, and profile completeness in identifying automated accounts (Cresci et al., 2015; Aljabri et al., 2023). Conversely, the poor performance of visual features suggests that, at least in the Instagram context, profile pictures are less dependable indicators of authenticity compared to numerical or behavioral metrics. This finding confirms earlier research by Akyon & Kalfaoglu (2019), who also observed the limited contribution of visual features to classification performance.

Textual features, like comments later added to the semi-supervised model, are usually reliable indicators for detecting synthetic online engagement, as several studies on X have demonstrated.

The use of multimodal learning adds value. While metrics dominate the feature importance rankings, the integration of visual features in fusion models slightly improved robustness across classifiers. This supports the argument made by Baltrušaitis et al. (2018) that multimodality can help compensate for missing or noisy information in individual modalities. In semi-supervised experiments, this became even more evident, as visual features contributed up to 30% of the overall feature importance in LightGBM models, highlighting their complementary role even if they are not strong predictors on their own. Additionally, features extracted from textual data could contribute up to 19% in the self-learning model used with the LightGBM classifier.

The approach of implementing the early fusion strategy was effective, as the classification results in terms of accuracy produced the best outcomes and laid the groundwork for future development of semi-supervised models.

As often stressed in this thesis semi-supervised models are not that common in the objective of classifying social media bots. As in any other form of applying ML on the classification problem X has significantly most studies around semi-supervised algorithms which can be seen in the work of Zeng et al. (2021). They classified bots on X with a similar strategy of training base classifiers and then reusing confident labels of the self-learning model to be attached to the training data.

5.2 Limitations

This thesis sets out to advance the understanding of bot detection on social media with a focus on Instagram and a multimodal approach. While meaningful contributions were made, a deployment of the final model should not be considered, as the results are not generalizable. Several constraints restrict the confidential implementation of such a system.

The first limitation concerns the size of the labeled dataset. Compared to other studies, the number of manually labeled accounts was rather small, which constrained the robustness and generalizability of the training data. The acquisition of unlabeled data in the form of comments, profile pictures, and profile metrics was more successful, but expanding this dataset further could strengthen the empirical foundation of the models.

A further limitation lies in the scope of the collected comments. Only original comments were included, while replies and non-textual interactions such as GIFs were excluded. As bots often engage through such media, these instances of automated behavior may have been missed. In addition, the removal of duplicate entries eliminated the opportunity to analyze posting frequency per user, which could have served as an indicator of automated activity.

The multilingual nature of the dataset also posed a challenge, making sentiment analysis difficult. Such analysis might have provided an additional perspective on genuine versus synthetic behavior. Similarly, while features from profile pictures were extracted, they did not significantly improve classification performance, which limits the contribution of the visual modality in this study.

Finally, while several classifiers with different hyperparameters were tested, performance improvements remained limited. Time and resource constraints prevented a more extensive manual labeling process or advanced annotation strategies such as cross-labeling, which would have helped to enlarge and balance the dataset.

5.3 Future Research

As automated behavior on social media is expected to become increasingly sophisticated, further research on this topic is essential. The scope should not remain limited to platforms such as X, Facebook, and Instagram but should also include TikTok, which has grown rapidly in recent years and attracted a worldwide user base. It would be valuable to investigate the extent of automated behavior on TikTok, particularly since the platform does not promote public discussions to the same degree as X but relies heavily on comment sections, which could also be targeted by bots.

As outlined in Chapter 4.8, future studies could aim to capture a more complete picture of user interactions by including replies to comments and incorporating non-textual elements such as GIFs. These interaction types might contain meaningful signals of automated behavior that are currently overlooked.

Another direction for future research is addressing the scarcity of labeled datasets. One possibility would be to collaborate directly with Instagram or similar platforms to gain access to verified bot accounts or previously removed accounts that violated platform policies. Such datasets would provide a more reliable foundation for model training and evaluation.

Additionally, sentiment analysis of comments represents an unexplored area in Instagram-focused research. Conducting sentiment analysis across different languages could enrich the detection process by highlighting emotional or behavioral differences between genuine users and bots.

From a methodological perspective, future work could move beyond traditional classifiers and explore more advanced neural network architectures. Neural approaches could be used to create richer multimodal representations that integrate text, profile metrics, and images. In this context, Generative Adversarial Networks (GANs) may also prove useful to alleviate the lack of labeled data by generating synthetic but realistic examples for model training.

When it comes to visual features, further research could consider implementing reverse image search functions to identify whether suspicious accounts rely on stock images or steal pictures from other public profiles. Previous research has shown that such approaches can add valuable information to the classification process, though they often require costly access to external APIs. If implemented, this type of feature could significantly enhance multimodal models.

Overall, future studies should combine improved data availability, novel feature types, and advanced ML techniques to better capture the complexity of automated online behavior.

6 CONCLUSION

This thesis addressed several research goal regarding Instagram bot detection using multimodal machine learning approaches. One of them was the development of a multimodal ML algorithm able to classify between genuine and synthetic behaviour. The multimodal framework achieved measurable improvements over single-modality approaches. While profile metrics alone reached 97% accuracy, the integration of visual and textual features through early fusion increased accuracy to 98%.

The semi-supervised approach successfully expanded the training dataset from 634 to 9,562 labeled instances through iterative self-training. The confidence-based selection (99% threshold) maintained label quality while achieving practical scale. The method labeled 89.3% of the unlabeled dataset with high confidence, demonstrating that semi-supervised learning can effectively operate in low-labeled-data environments common in Instagram research.

Profile metrics emerged as the most predictive features, with bioLength, postsCount, and follower ratios consistently ranking highest in importance. Visual features, while individually weak (62-69% accuracy), provided valuable complementary information in multimodal models. Textual features from comments contributed meaningfully despite

language diversity challenges. The findings suggest that behavioral and structural indicators remain more reliable than visual or linguistic patterns for Instagram bot detection.

This work contributes to both academic knowledge and practical application in several ways. It presents the first comprehensive multimodal study of Instagram bot detection that integrates profile metrics, visual features, and textual analysis into a single framework. By doing so, it validates the applicability of Georges' identity layer model as a theoretical foundation for understanding and analyzing synthetic behavior on social media platforms. Furthermore, the work demonstrates the effectiveness of semi-supervised learning in contexts where only limited labeled data is available, which reflects a common challenge in social media research. Finally, the thesis advances methodological practice by showing how the DSR approach can be combined with the CRISP-DM process model to provide a structured and rigorous framework for social media analytics research.

The DSR framework successfully guided this thesis from problem identification through artifact evaluation. The knowledge base, established through literature review, provided crucial theoretical foundations while the environment justified practical relevance.The iterative artifact development process demonstrated DSR's core principles. Multiple rounds of model training and evaluation, informed by both theoretical knowledge and practical constraints, produced progressively refined solutions. The evaluation methods validated the artifact against both knowledge base criteria and environmental needs. The balanced consideration of scientific rigor and practical utility exemplifies successful DSR implementation.

LIST OF KEY LITERATURE

1. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. (2016). The rise of social bots. *Communications Of The ACM, 59*(7), 96–104. https://doi.org/10.1145/2818717
2. Imperva. (2023). *Imperva bad bot report*. https://www.imperva.com/resources/reports/2023-Imperva-Bad-Bot-Report.pdf
3. Hevner, A. R., March, S. T., Park, J, & Sudha R. (2004). *Design science in Information Systems Research*. https://www.researchgate.net/publication/201168946_Design_Science_in_Information_Systems_Research
4. Muzumdar, P., Cheemalapati, S., RamiReddy, S. R., Singh, K., Kurian, G., & Muley, A. (2025). The Dead Internet Theory: A survey on Artificial interactions and the future of social media. *Asian Journal of Research in Computer Science, 18*(1), 67–73. https://doi.org/10.9734/ajrcos/2025/v18i1549

5. Narayanan, A. (2023). *Understanding social media recommendation algorithms*. Knight First Amendment Institute. https://s3.amazonaws.com/kfai-documents/documents/4a9279c458/Narayanan---Understanding-Social-Media-Recommendation-Algorithms_1-7.pdf
6. Ng, L. H. X. & Carley, K. M. (2025). *A global comparison of social media bot and human characteristics*. Scientific Reports, 15:10973. https://doi.org/10.1038/s41598-025-96372-1
7. Wirth, R., Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining*. https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

REFERENCE LIST

1. Aiello, L. M., Deplano, M., Schifanella, R., & Ruffo, G. (2021). People are strange when you're a stranger: Impact and influence of bots on social networks. *Proceedings of the International AAAI Conference on Web and Social Media, 6*(1), 10–17. https://doi.org/10.1609/icwsm.v6i1.14236
2. Akyon, C. F., Kalfaoglu, E., M., (2019). *Instagram Fake and Automated Account Detection*. IEEE Conference Publication. https://ieeexplore.ieee.org/abstract/document/8946437
3. Aljabri, M., Zagrouba, R., Shaahid, A., Alnasser, F., Saleh, A., & Alomari, D. M. (2023). Machine learning-based social media bot detection: a comprehensive literature review. *Social Network Analysis and Mining, 13*, 20. https://doi.org/10.1007/s13278-022-01020-5
4. Andriotis, P. & Takasu, A., (2019). *Emotional Bots: Content-based Spammer Detection on Social Media*. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8630760
5. Arjel, G., Chowdhury, A., Tiwari, M. (2025). Dead Internet Theory and how Generative AI has contributed to realizing it. *International Research Journal of Education and Technology,.* *7*(2). https://www.irjweb.com/Dead%20Internet%20Theory%20and%20how%20Generative%20AI%20has%20contributed%20to%20realizing%20it.pdf
6. Backstrom, L. S. (2013). *News Feed FYI: A window into News Feed*. Facebook Business. https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed
7. Baltrusaitis, T., Ahuja, C., Morency, L. P. (2018). Multimodal Machine Learning: A Survey and Taxonomy. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 41*, 423. https://doi.org/10.1109/TPAMI.2018.2798607
8. Bayer, J. B., Triệu, P., & Ellison, N. B. (2020). Social media elements, ecologies, and effects. *Annual Review of Psychology*, *71:*471–497.

https://www.annualreviews.org/docserver/fulltext/psych/71/1/annurev-psych-010419-050944.pdf?expires=1747832033&id=id&accname=guest&checksum=0AFF3FF58D0818A00F3FF253EA013CF8
9. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5
10. Berners- Lee ;T. (1998): T*he World Wide Web: A very short personal history*. https://www.w3.org/People/Berners-Lee/ShortHistory.html
11. Bessi, A., & Ferrara, E. (2016). *Social bots distort the 2016 US presidential election online discussion*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2982233
12. Beskow, D., Carley, K., (2018). *Bot-hunter: A Tiered Approach to Detecting & Characterizing Automated Activity on Twitter*. (Conference Paper). https://www.researchgate.net/publication/326606376
13. Blumauer, A., & Pellegrini, T. (2008). *Semantic Web Revisited – Eine kurze Einführung in das Social Semantic Web*. X.media.press (pp. 3–22). https://doi.org/10.1007/978-3-540-72216-8_1
14. Boyd, D. M., & Ellison, N. B. (2010). Social network sites: definition, history, and scholarship. *IEEE Engineering Management Review, 38*(3), 16–31. https://doi.org/10.1109/emr.2010.5559139
15. Bucher, T. (2018). *If . . . then : algorithmic power and politics*. Oxford University Press.
16. Burgess, J., & Green, J. B. (2010). *YouTube: Online video and participatory culture*. https://www.researchgate.net/publication/27478259_YouTube_Online_Video_and_Participatory_Culture
17. Burkov, A. (2019.). *The Hundred-Page Machine Learning Book*.
18. Buzzoid. (n. d.). *Buy Instagram Followers with instant delivery*. https://buzzoid.com/buy-instagram-followers/
19. Castells, M. (2009). *The rise of the network society*.
20. Chen, J. V., Nguyen, T., Jaroenwattananon, J. (2021). WHAT DRIVES USER ENGAGEMENT BEHAVIOR IN a CORPORATE SNS ACCOUNT: THE ROLE OF INSTAGRAM FEATURES. *Journal of Electronic Commerce Research, 22*(3), 199–200. http://www.jecr.org/sites/default/files/2021vol22no3_Paper2.pdf
21. Choudhury, N. (2014). World Wide Web and Its Journey from Web 1.0 to Web 4.0. *IJCSIT International Journal Of Computer Science And Information Technologies, 5–6*, 8096–8100. https://ijcsit.com/docs/Volume%205/vol5issue06/ijcsit20140506265.pdf
22. Carr, C. T., & Hayes, R. A. (2015). Social Media: defining, developing, and divining. *Atlantic Journal of Communication, 23(1),* 46–65. https://doi.org/10.1080/15456870.2015.972282
23. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2015). Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems, 80*, 56–71. https://doi.org/10.1016/j.dss.2015.09.003

24. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A. & Tesconi, M. (2017). The Paradigm-Shift of Social Spambots. *International World Wide Web Conference Committee (IW3C2),* 963–972. https://doi.org/10.1145/3041021.3055135
25. Cresci, S. (2020). A Decade of Social Bot Detection. *COMMUNICATIONS OF THE ACM, 63*(10), 73–74. https://doi.org/10.1145/3409116
26. Dewan, P., Kumaraguru, P. (2015). *Detecting Malicious Content on Facebook.* Association For The Advancement Of Artificial Intelligence.
27. Dimitriadis, I., Georgiou, K., & Vakali, A. (2021). Social Botomics: A Systematic Ensemble ML approach for Explainable and Multi-Class Bot Detection. *Applied Sciences, 11*(21), 9857. https://doi.org/10.3390/app11219857
28. Di Gangi, P. M. & Wasko, M. (2016). Social Media Engagement Theory: Exploring the Influence of User Engagement on Social Media Usage. *Journal Of Organizational And End User Computing, 28*(2), 53–73. https://doi.org/10.4018/JOEUC.2016040104
29. Di Placido, D. (2024). *The dead internet theory, explained*. Forbes. https://www.forbes.com/sites/danidiplacido/2024/01/16/the-dead-internet-theory-explained/
30. Facebook (2012). *'Facebook to acquire Instagram'. Facebook Newsroom*. https://newsroom.fb.com/news/2012/04/facebook-to-acquireinstagram/
31. Talavera, O., Tran, V. (2019). *Social media bots and stock markets*. https://doi.org/10.13140/RG.2.2.32949.32482
32. Fang, W., Nie, C. (2024) Social media use, social bot literacy, perceived threats from bots, and perceived bot control: a moderated-mediation model. *Behaviour & Information Technology, 43*(13), 3271-3287
33. Ferrara, E., Jr., Varol, O., Davis, C., Menczer, F. & Flammini, A. (2016a). Today's social bots: sophisticated tools for social media manipulation. *Communications Of The ACM, . 59*(7), 96–99. https://dl.acm.org/doi/pdf/10.1145/2818717
34. Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. (2016). The rise of social bots. *Communications Of The ACM, 59*(7), 96–104. https://doi.org/10.1145/2818717
35. GeeksforGeeks. (2025). *LightGBM (Light Gradient Boosting Machine)*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/lightgbm-light-gradient-boosting-machine/
36. Georges, F. (2011). *L'identité numérique sous emprise culturelle:De l'expression de soi à sa standardisation*. https://shs.cairn.info/revue-les-cahiers-du-numerique-2011-1-page-31?lang=fr
37. Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.
38. Goodrow, C. (2025). *On YouTube's recommendation system*. blog.youtube. https://blog.youtube/inside-youtube/on-youtubes-recommendation-system/
39. Goyal, B., Gill, N. S., Gulia, P., Prakash, O., Priyadarshini, I., Sharma, R., Obaid, A. J. & Yadav, K. (2023). Detection of Fake Accounts on Social Media Using Multimodal Data With Deep Learning. *IEEE Transactions On Computational Social Systems*, 1–12. https://doi.org/10.1109/tcss.2023.3296837

40. Grosser, B. (2014). *What do metrics want? How quantification prescribes social interaction on Facebook*. http://computationalculture.net/what-do-metrics-want/
41. Habib, A. K. M. R. R., Akpan, E. E., Ghosh, B. & Dutta, I. K. (2024). Techniques to Detect Fake Profiles on Social Media Using the New Age Algorithms - A Survey. *IEEE 12th Annual Computing And Communication Workshop And Conference (CCWC)*, 0329–0335. https://doi.org/10.1109/ccwc60891.2024.10427620
42. Hagen, L., Neely, S., Keller, T. E., Scharf, R., & Vasquez, F. E. (2020). Rise of the Machines? Examining the influence of social bots on a political discussion network. *Social Science Computer Review, 40*(2), 264–287. https://doi.org/10.1177/0894439320908190
43. Hevner, A. R., March, S. T., Park, J, & Sudha R. (2004). *Design Science in Information Systems Research*. https://www.researchgate.net/publication/201168946_Design_Science_in_Information_Systems_Research
44. Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What we Instagram: A first analysis of Instagram photo content and user types. *Proceedings of the International AAAI Conference on Web and Social Media, 8*(1), 595–598. https://doi.org/10.1609/icwsm.v8i1.14578
45. Imperva. (2023). *Imperva bad bot report*. https://www.imperva.com/resources/reports/2023-Imperva-Bad-Bot-Report.pdf
46. Kaplan, A. M., & Haenlein, M. (2009). The fairyland of Second Life: Virtual social worlds and how to use them. *Business Horizons, 52*(6), 563–572. https://doi.org/10.1016/j.bushor.2009.07.002
47. Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003
48. Kilian, T., Langner, S. (2010). *Social-Media-Kommunikation: Web 2.0-Dienste aktiv nutzen*. Online-Kommunikation. Gabler Verlag, Wiesbaden. https://doi.org/10.1007/978-3-8349-8897-3_12
49. Kollanyi, B., Howard, P. N., Woolley, S. C. (2016). Bots and Automation over Twitter during the U.S. Election. *COMPROP DATA MEMO 2016.4.* https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/12/2016/11/Data-Memo-US-Election.pdf.
50. Kuhn, M. & Johnson, K. (2018). *Applied Predictive Modeling*. Springer.
51. Lammenett, E. (2024). *Social-Media-Marketing, Web 2.0 und Co.: Einordnung und Hintergrundwissen sowie Details zu den wichtigsten Themen*. Praxiswissen Online-Marketing. Springer Gabler, Wiesbaden. https://doi.org/10.1007/978-3-658-43610-0_17
52. Leaver, T., Highfield, T. & Abidin, C. (2020). *Instagram: Visual Social Media cultures*. Polity.
53. Lee, K., Eoff, B.D., and Caverlee, J. (2011). Seven months with the devils: A long-term study of content polluters on Twitter. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 185–192.

54. Mariani, R. (2023). The dead internet to come. *The New Atlantis, 73*, 34–42. https://www.jstor.org/stable/27244117
55. McDonough, D. J., Helgeson, M. A., Liu, W., & Gao, Z. (2021). Effects of a remote, YouTube-delivered exercise intervention on young adults' physical activity, sedentary behavior, and sleep during the COVID-19 pandemic: Randomized controlled trial. *Journal of Sport and Health Science/Journal of Sport and Health Science, 11*(2), 145–156. https://doi.org/10.1016/j.jshs.2021.07.009
56. Meshram, E. P., Bhambulkar, R., Pokale, P., Kharbikar, K., & Awachat, A. (2021). Automatic detection of fake profile using machine learning on Instagram. *International Journal of Scientific Research in Science and Technology*, 117–127. https://doi.org/10.32628/ijsrst218330
57. Meta for business. (n.d.). *Meta for Business*. https://www.facebook.com/business/news/News-Feed-FYI-A-Window-Into-News-Feed
58. Meta. (n.d.a). Brand Resource Center. https://about.meta.com/de/brand/resources/meta/our-trademarks/
59. Meta. (n.d.b). Brand Transparency Center. https://transparency.meta.com/features/explaining-ranking/fb-feed-recommendations/
60. Meta. (n.d.c). Brand Transparency Center. https://transparency.meta.com/features/explaining-ranking/ig-feed-recommendations/
61. Meta (n.d.d) *How we are using facial recognition technology on images to keep people safe on our platforms* | Meta-Hilfebereich. https://www.meta.com/de-de/help/policies/1958769918293906/
62. Meta Platforms Ireland Limited. (2022). *NetzDG Transparenzbericht*. https://about.fb.com/de/wp-content/uploads/sites/10/2022/07/Facebook-NetzDG-Transparenzbericht-July-2022.pdf
63. Metcalfe, B. (2013). Metcalfe's Law after 40 Years of Ethernet. *Computer*, *46*(12), 26–31. https://doi.org/10.1109/mc.2013.374
64. Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Science/Engineering/Math
65. Mosseri, A. (2021). *Shedding more light on how instagram works*. Instagram. https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works
66. Murphy, K. P. (2012). *Machine learning: A Probabilistic Perspective*. MIT Press.
67. Muzumdar, P., Cheemalapati, S., RamiReddy, S. R., Singh, K., Kurian, G., & Muley, A. (2025). The Dead Internet Theory: A survey on Artificial interactions and the future of social media. *Asian Journal of Research in Computer Science, 18*(1), 67–73. https://doi.org/10.9734/ajrcos/2025/v18i1549
68. Narayanan, A. & Knight First Amendment Institute. (n.d.). *Understanding social media recommendation algorithms*. In: *Knight First Amendment Institute*. https://s3.amazonaws.com/kfai-documents/documents/4a9279c458/Narayanan---Understanding-Social-Media-Recommendation-Algorithms_1-7.pdf
69. Ng, L. H. X., Robertson, D. C., & Carley, K. M. (2016). *Cyborgs for strategic communication on social media*. In: SAGE. https://doi.org/10.1177/ToBeAssigned

70. Ng, L. H. X. & Carley, K. M. (2025). A global comparison of social media bot and human characteristics. *Scientific Reports*, *15*(1), 10973. https://doi.org/10.1038/s41598-025-96372-1
71. O'Reilly, T. (2005). *What is Web 2.0.* O'Reilly Media. https://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html?page=1
72. PwC. (2017). *Social Bots: Gefahr für die Demokratie?* https://www.pwc.de/de/technologie-medien-und-telekommunikation/social-bots-gefahr-fuer-die-demokratie.pdf
73. Radivojevic, K., McAleer, C., & Center for Research Computing. (2024). *Social Media Bot Policies: Evaluating passive and active enforcement*. University of Notre Dame. https://arxiv.org/pdf/2409.18931
74. Ray, S., Kim, S. S. & Morris, J. G. (2014). The Central Role of Engagement in Online Communities. *Information Systems Research, 25*(3) *528–546*. https://doi.org/10.1287/isre.2014.0525
75. Reda, E. Y. (2024). Exploring the components of digital identity on social networks sites: identifier, post, profile photo, and selfie. *European Scientific Journal ESJ, 20*(1), 1. https://doi.org/10.19044/esj.2024.v20n1p1
76. Sander, T., Teh, P. -L. (2014) Determining the indicators of social capital theory to social network sites, *3rd International Conference on User Science and Engineering (i-USEr)*, Shah Alam, Malaysia, 264-268, 10.1109/IUSER.2014.7002714
77. Santia, G. C., Mujib, M. I., Williams, J. R. (2019) Detecting Social Bots on Facebook in an Information Veracity Context. *Proceedings Of The Thirteenth International AAAI Conference On Web And Social Media (ICWSM 2019)*
78. Schultz, A. P., Piepgrass, B., Weng, C.-C., Ferrante, D., Verma, D., Martinazzi, P., Alison, T., & Mao, Z. (2014). *Methods and systems for determining use and content of PYMK based on value model*. In: *Patent Application Publication*. https://patentimages.storage.googleapis.com/36/6c/e3/e1b8090e95ce9b/US20140114774A1.pdf
79. Sengar, S. S., Kumar, S., Raina, P., & Mahaliyan, M. (2020). *Bot detection in social networks based on multilayered deep learning approach*. https://www.researchgate.net/publication/347439756_Bot_Detection_in_Social_Networks_Based_on_Multilayered_Deep_Learning_Approach
80. Shao, C., Ciampaglia, G. L., Varol, O., Yang, K., Flammini, A. & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications, 9*(1). https://doi.org/10.1038/s41467-018-06930-7
81. Sheehan, A. (n.d.). *The Dead Internet Theory*. https://dailyfreepress.com/10/21/21/182334/the-dead-internet-theory/
82. Shokeen, J., & Rana, C. (2019). A study on features of social recommender systems. *Artificial Intelligence Review, 53*(2), 965–988. https://doi.org/10.1007/s10462-019-09684-w
83. Shih, C. C. (2011). *The Facebook era: Tapping Online Social Networks to Market,* Sell, and Innovate. Addison-Wesley Professional.

84. Sleeman, W. C., IV, Kapoor, R., Ghosh, P. (2022). Multimodal Classification: Current Landscape, Taxonomy and Future Directions. *ACM Comput. Surv., 55*(7), 150–150. https://doi.org/10.1145/3543848
85. Statista. (2024). *Social media usage worldwide*
86. Subramanian, K. (2017). Influence of Social Media in Interpersonal Communication. *INTERNATIONAL JOURNAL OF SCIENTIFIC PROGRESS AND RESEARCH (IJSPR)*, *02*, 70–70. https://www.ijspr.com/citations/v38n2/IJSPR_3802_2069.pdf
87. Sykora, M. (2017b). Web 1.0 to Web 2.0: an observational study and empirical evidence for the historical r(evolution) of the social web. *International Journal Of Web Engineering And Technology, 12*(1), 70. https://doi.org/10.1504/ijwet.2017.084024
88. Taylor, D. G., Lewin, J. E., & Strutton, D. (2011). Friends, fans, and followers: Do ads work on social networks? *Journal of Advertising Research, 51(*1), 258–275. https://doi.org/10.2501/jar-51-1-258-275
89. Van Dijck, J. (2013). *The Culture of Connectivity*. Oxford University Press.
90. Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). *Online Human-Bot Interactions: Detection, Estimation, and Characterization*. https://arxiv.org/abs/1703.03107
91. Wang, G., Konolige, T., Wilson, C., Wang, X., Zheng, H., Zhao, B. Y., University of California, Santa Barbara, Northeastern University & Renren Inc. (2013). You Are How You Click: Clickstream Analysis for Sybil Detection. *Proceedings Of The 22nd USENIX Security Symposium*, 241. https://www.usenix.org/system/files/conference/usenixsecurity13/sec13-paper_wang_0.pdf
92. Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2012, May 17). *Social Turing Tests: Crowdsourcing Sybil Detection*. https://arxiv.org/abs/1205.3856
93. Walter, Y. (2024). Artificial influencers and the dead internet theory. *AI & Society*, *40*(1), 239–240. https://doi.org/10.1007/s00146-023-01857-0
94. Weinberg, T. & Pahrmann, C. (2014). *Social Media Marketing - Strategien für Twitter, Facebook & Co*.
95. Wirth, R., Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining*. https://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf
96. Wu, B., Liu, L., Yang, Y., Zheng, K. & Wang, X. (2020). Using Improved Conditional Generative Adversarial Networks to Detect Social Bots on Twitter. *IEEE Access*, *8*, 36664–36680. https://doi.org/10.1109/access.2020.2975630
97. Zarei, K., Farahbakhsh, R., Crespi N. (2019) Typification of Impersonated Accounts on Instagram. *IEEE 38th International Performance Computing and Communications Conference (IPCCC)* 1-6, https://doi.org/10.1109/ipccc47392.2019.8958763
98. Zeng, Z., Li, T., Sun, S., Sun, J., & Yin, J. (2021). A novel semi-supervised self-training method based on resampling for Twitter fake account identification. *Data Technologies and Applications, 56*(3), 409-428

99. Zhang, X., Liu, J., & Xu, Z. (2015). Tencent and Facebook data validate Metcalfe's law. *Journal of Computer Science and Technology, 30*(2), 246–251. https://doi.org/10.1007/s11390-015-1518-1

APPENDICES

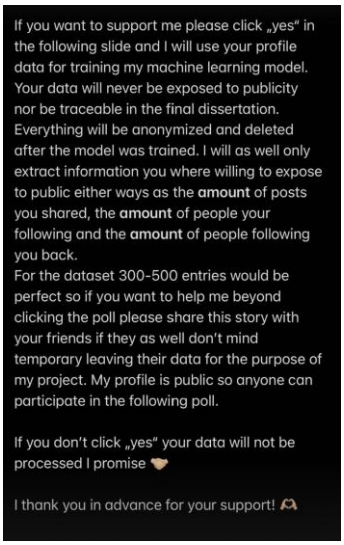
Appendix 1: Existing work sorted by features

Table 9: Existing work sorted by features

Class	**Features**
Network	Extraction of relationships, Contributions, Mentioning and hashtag mapping, Percentage of friends with profile descriptions (Varol et al., 2017, p. 2-3; Dimitriadis et al., 2021, p. 11)
User	Number of friends, Number of followers, Number of posts, Self-description via bios, Username length, Account age, Default profile pic, Verified profile, Active status/story (Akyon & Kalfaoglu, 2019, p. 2; Goyal et al., 2023, p. 5; Meshram et al., 2021, p. 121; Varol et al., 2017, p. 2-3)
Friends	Amount of followers, Amount of following, Ratio of followers/following, (Akyon & Kalfaoglu, 2019, p. 2; Meshram et al., 2021, p. 121)
Timing	Average publishing of posts, Response time (Varol et al., 2017, p. 2-3; Beskow & Carley, 2018, p. 6)
Content	Length of post description, Type of words used (verbs, nouns, adjectives), Ratio of hashtags used to total posts, Ratio of mentions (@) used to total posts (Varol et al., 2017, p. 2-3; Andriotis, P. & Takasu, A., 2019, p. 3)
Sentiment	Dominance, Happiness, Polarization, Emotions through emoticons, Similarity of contents (Varol et al., 2017, p. 2-3; Wu et al., 2020, p. 36669)

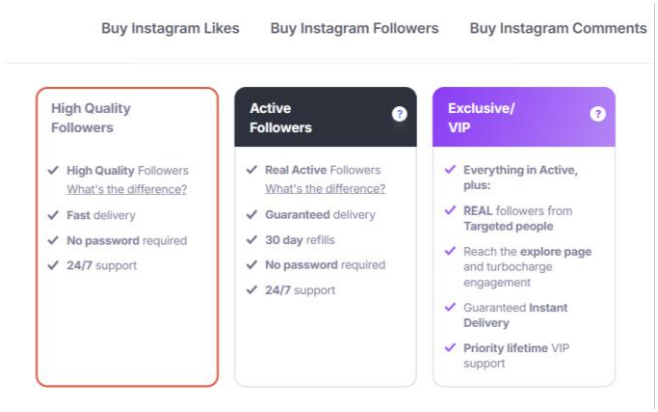
Source: Own work based on Ferrera (2017).

Appendix 2: Survey for profile observation of genuine accounts



Source: Own work

Appendix 3: Buying options on Buzzoid.com



Source: Buzzoid. (n. d.)

Appendix 4: Scrapped Instagram comment sections

Table 10: Scrapped Instagram comment sections

Meme pages	**Financial pages**	**Politicians**
@9gag	@bloombergbusiness	@bundeskanzler
@daquan	@chime	@christianlindner
@epicfunnypage	@financialtimes	@claudia_shein
@memezar	@marketwatch	@donaldtusk
@trashcanpaul	@personalfinanceclub	@emmanuelmacron
	@thefinancialdiet	@jdvance
	@traderepublic	@justinpjtrudeau
	@traderepublic_de	@minpresrutte
	@wealth	@rterdogan
	@yahoofinance	@shinozabe
		@theresamay
		@ursulavonderleyen

Source: Own work

Appendix 5: Anonymized data snippet of comment features

text_div	heading	Length character	# words	# mentions	# hastags	URL included?	Emojis?	Response time in minutes
Brazil is sovereign! ð`	[redacted]	29	4	0	0	no	yes	87,8
Epstein files		13	2	0	0	no	no	3,749999997
What about the Espt		29	5	0	0	no	no	10,49999999
Heâ€™s making it up		37	8	0	0	no	no	2,633333331
Majority of your big c		127	23	0	0	no	no	120,1833333
He'll do anything to a		51	9	0	0	no	yes	8,999999998
Welcome to India Ch		39	6	0	0	no	no	64,63333333
The world needs to tu		103	22	0	0	no	no	21,16666666
Go BRICS âœï¸		15	3	0	0	no	no	196,6666667
Apple just moved a lo		94	16	0	0	no	yes	97,86666666
Since most of the IPh		101	22	0	0	no	no	1524,683333
How much will this ir		59	10	1	0	no	no	612,1333333
Epstein		7	1	0	0	no	no	508,3166667
Great .. punish us fur		55	11	0	0	no	no	998,6166667
For 48 hours only… It		40	7	0	0	no	no	111,25
India will not bow do		23	5	0	0	no	no	8,099999997
The man truly doesn'		86	14	0	0	no	no	347,55
I thought it we were c		40	9	0	0	no	no	19,1
From India. Us peopl		63	12	0	0	no	yes	1,033333334
But don't make an ex		48	7	0	0	no	no	42,54999999
25% tax on American		29	5	0	0	no	no	2,783333334
De-dollarization on B		36	5	0	0	no	yes	1110,066667
He is literally diminis		82	15	0	0	no	yes	5,666666661
ðŸ¨,ðŸ‘ prices going		42	8	0	0	no	yes	70,16666666
Donald Duck has dar		55	8	0	0	no	no	47,93333333

Appendix 6: Engineering of profile picture features

```python
# Feature array
data = []

# Load ResNet-Modell
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
resnet = models.resnet50(pretrained=True)
resnet = torch.nn.Sequential(*(list(resnet.children())[:-1]))  # letzte Schicht entfernen
resnet.eval().to(device)

# Transformations for ResNet
transform_resnet = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor(),
    transforms.Normalize(mean=[0.485, 0.456, 0.406],
                         std=[0.229, 0.224, 0.225])
])

# Extraction of features
def get_color_features(img):
    """Average color"""
    hsv = cv2.cvtColor(img, cv2.COLOR_BGR2HSV)
    mean_color = hsv.mean(axis=(0, 1))  # Hue, Saturation, Value
    return mean_color.tolist()

def get_sharpness(img):
    """Detection of sharpness"""
    gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
    return cv2.Laplacian(gray, cv2.CV_64F).var()

def has_face(img):
    """Face recognition"""
    gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
    face_cascade = cv2.CascadeClassifier(cv2.data.haarcascades + "haarcascade_frontalface_default.xml")
    faces = face_cascade.detectMultiScale(gray, 1.1, 4)
    return 1 if len(faces) > 0 else 0

def get_resnet_features(pil_img):
    """ResNet-Embedding (2048-Dim)"""
    tensor = transform_resnet(pil_img).unsqueeze(0).to(device)
    with torch.no_grad():
        features = resnet(tensor)
    return features.squeeze().cpu().numpy()
```

Source: Own work

Appendix 7: Baseline Model early fusion with Random Forest Classifier

```
#Searching for rows where null-values exist to delete them
missing_report = df.isnull().sum().sort_values(ascending=False)
print(missing_report)

rows_with_zeros = df[df.isnull().any(axis=1)]

print(f"{len(rows_with_zeros)} rows with a minimum of one null-value.")
print(rows_with_zeros)
```

✓ 0.0s

```
username              0
mean_hue              0
bio length            0
numbers in username    0
length username        0
externalUrl           0
followersCount        0
followsCount          0
highlightReelCount    0
igtvVideoCount        0
isBusinessAccount     0
postsCount            0
private               0
face_detected         0
sharpness             0
mean_value            0
mean_saturation       0
Bot?                  0
dtype: int64
0 rows with a minimum of one null-value.
```

```
#Transform textual boolean values into numerical values
bool_spalten = ["private", "joinedRecently", "isBusinessAccount", "hasChannel", "face_detected", "externalUrl", "hasBio"]

df[bool_spalten] = df[bool_spalten].replace({True: 1, False: 0})
```

```
#Dropping the columns where there is only either values of 0 or 1
df = df.drop(columns=['hasChannel', 'joinedRecently', "hasBio"])
```

```python
#Drop the label fot the dataset used and seperately save it
features_df = df.drop(columns=['Bot?'])
target = df['Bot?']
```

```python
#Import the train and test split and determine X and y for the ML Model, as well as dropping the usernames
from sklearn.model_selection import train_test_split

X = features_df.drop(columns="filename")
y = df['Bot?']
```

```python
#Implementing the train and test splits of 80 to 20 and displaying the shapes
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, shuffle=True, random_state=42)
print(f"The shape of the training data is {X_train.shape}")
print(f"The shape of the test data is {X_test.shape}")
```

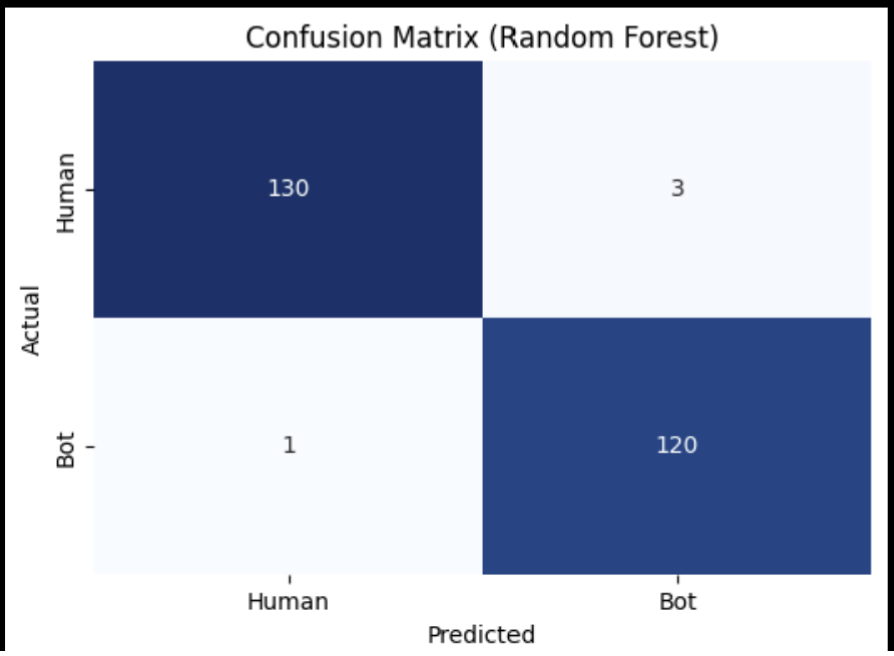
```
The shape of the training data is (380, 16)
The shape of the test data is (254, 16)
```

```python
#Import the Standardscaler
from sklearn.preprocessing import StandardScaler
from sklearn.compose import ColumnTransformer

# Identify numeric columns
numeric_features = X_train.select_dtypes(include=['int64', 'float64']).columns

# Scale only numeric columns
preprocessor = ColumnTransformer([('scaler', StandardScaler(), numeric_features)], remainder='passthrough')
X_train_scaled = preprocessor.fit_transform(X_train)
X_test_scaled = preprocessor.transform(X_test)
```

```python
# Importing necessary libraries
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score, precision_score, recall_score, f1_score, confusion_matrix, roc_auc_score, roc_curve
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import pickle

# Random Forest with hyperparameters
rf = RandomForestClassifier(
    random_state=42,
    n_estimators=100,        # Amount of trees
    class_weight='balanced', # Weigthing of the classes
    max_depth=10,            # Maximum depth of trees
    min_samples_split=5,     # Minimum samples for the split of a knot
    min_samples_leaf=2,      # Minimum samples in a leaf
    max_features='sqrt',     # Amount features per split
    bootstrap=True,          # Bootstrap-Sampling for every tree
)

# Training of the Random Forest Modell
rf.fit(X_train_scaled, y_train)

# Predictions
y_pred = rf.predict(X_test_scaled)

# Calculate the metrics for evaluation
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

# ROC Curve
y_prob = rf.predict_proba(X_test_scaled)[:, 1]
fpr, tpr, thresholds = roc_curve(y_test, y_prob)
roc_auc = roc_auc_score(y_test, y_prob)

# Printing out the results
print("Classification Report:")
print(classification_report(y_test, y_pred))
```

```python
# Feature Importance (Random Forest spezifisch)
feature_importance = rf.feature_importances_
print(f"\nTop 5 most important features:")
if hasattr(X_train_scaled, 'columns'):
    feature_names = X_train_scaled.columns
else:
    feature_names = [f'Feature_{i}' for i in range(len(feature_importance))]

importance_df = pd.DataFrame({
    'Feature': feature_names,
    'Importance': feature_importance
}).sort_values('Importance', ascending=False)

print(importance_df.head())

# Confusionsmatrix
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(6,4))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", cbar=False,
            xticklabels=['Human', 'Bot'],
            yticklabels=['Human', 'Bot'])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix (Random Forest)")
plt.show()

# ROC-Curve
plt.figure(figsize=(6,4))
plt.plot(fpr, tpr, label=f'ROC Curve (AUC = {roc_auc:.2f})')
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (Random Forest)')
plt.legend(loc="lower right")
plt.show()

# Feature importance Plot
plt.figure(figsize=(10, 6))
top_10_features = importance_df.head(10)
plt.barh(range(len(top_10_features)), top_10_features['Importance'])
plt.yticks(range(len(top_10_features)), top_10_features['Feature'])
plt.xlabel('Feature Importance')
plt.title('Top 10 Importance (Random Forest)')
plt.gca().invert_yaxis()
plt.tight_layout()
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.98      0.98       133
           1       0.98      0.99      0.98       121

    accuracy                           0.98       254
   macro avg       0.98      0.98      0.98       254
weighted avg       0.98      0.98      0.98       254


Detailed Metrics:
Accuracy:  0.9843
Precision: 0.9756
Recall:    0.9917
F1-Score:  0.9836
ROC-AUC:   0.9994
OOB Score: 0.9842

Top 5 wichtigste Features:
       Feature  Importance
11  Feature_11    0.394911
6    Feature_6    0.178927
15  Feature_15    0.175690
9    Feature_9    0.086203
14  Feature_14    0.044662
```



Source: Own work

Appendix 8: Snippet of the final model incorporating RF and LGBM for classification

```
# Model training
pipeline.fit(X_combined, y_combined)

# Predictions for unlabeled data
unlabeled_indices = range(len(X_labeled), len(X_combined))
X_unlabeled_for_pred = X_combined.iloc[unlabeled_indices]

predictions = pipeline.predict(X_unlabeled_for_pred)
probabilities = pipeline.predict_proba(X_unlabeled_for_pred)

# Filter results with 95% confidence
confidence_threshold = 0.95
high_confidence_mask = np.max(probabilities, axis=1) >= confidence_threshold

high_confidence_predictions = predictions[high_confidence_mask]
high_confidence_probabilities = probabilities[high_confidence_mask]

print(f"\nResults:")
print(f"Total unlabeled samples: {len(X_unlabeled)}")
print(f"Samples with ≥95% confidence: {np.sum(high_confidence_mask)}")
print(f"Confidence rate: {np.sum(high_confidence_mask)/len(X_unlabeled):.2%}")

# Create new labeled data
new_labeled_data = unlabeled_data.copy()
new_labeled_data['label'] = predictions
new_labeled_data['prediction_confidence'] = np.max(probabilities, axis=1)

# Transform unlabeled data for separate confidence calculations
X_unlabeled_transformed = pipeline.named_steps['preprocessor'].transform(X_unlabeled)

# Add separate confidence values for both models
new_labeled_data['rf_confidence'] = np.max(
    pipeline.named_steps['classifier'].classifier1.predict_proba(X_unlabeled_transformed),
    axis=1
)
new_labeled_data['lgb_confidence'] = np.max(
    pipeline.named_steps['classifier'].classifier2.predict_proba(X_unlabeled_transformed),
    axis=1
)

# Save only high-confidence samples
high_confidence_data = new_labeled_data[high_confidence_mask].copy()

# Save results
high_confidence_data.to_csv('new_labeled_data_high_confidence_CoTraining.csv', index=False)
new_labeled_data.to_csv('all_predictions_CoTraining.csv', index=False)
```

Source: Own work