

MSc Cyber Security Project – April 2025
Dissertation

Dr Anupam Mazumdar
MSc Cyber Security Project – April 2025

University of Essex (Online) | Computing Department
10th November 2025

Title

To explore the Ethics of a Data Science Company, Influencing Audiences through campaigns on the Social Media Platform, in the context of Facebook.

Name

Amrol Miah

Student No

12693968

Subject area

MSc in Cyber Security – University of Essex Online

e-Portfolio: <https://abmiah.github.io/eportfolio/index.html>

GitHub: <https://github.com/abmiah/eportfolio>

Coding Demo: <https://www.youtube.com/watch?v=aCb6T-xMF7w>

Content Page

1. Chapter 1: Introduction

(Overview, Aim, Objective, Research Questions, Scope, Rationale, Structure)

2. Chapter 2: Literature Review

(Background, Who is Cambridge Analytica, Cambridge Analytica Scandal Overview, Third-party Application, Influence and Ethics, Synthesis and Research Gaps)

3. Chapter 3: Methodology

(Introduction, Research Design, Data Source, Data Preparation, Analytical Methods, Ethical Considerations, Limitations, Summary)

4. Chapter 4: Results and Discussion

(Introduction, Descriptive Statistics, Cluster Analysis, Discussion, Mitigation Strategy, Summary)

5. Chapter 5: Conclusion

(Research Aim and Methods, Main Findings, Implications, Ethics, Limitations, Future Research, Final Reflection)

6. References

7. Bibliography

8. Appendix

Chapter 1:

1. Introduction

The term “Data” has become increasingly prevalent in the twenty-first century. When people consider data, it is often seen as a valuable resource, often compared to modern equivalents like “digital gold,” particularly for malicious actors. Cybercriminals who acquire data from institutions usually monetise these records by either selling them to other malicious actors or encrypting the victim organisations’ data and demanding ransom payments (Boyce et al., 2025). Consequently, various industries have instituted quantifiable security measures to protect user data from malicious threats that might exploit it.

However, the question raised is: what is data, and what significance does it have? According to the Cambridge Dictionary (2019), data refers to a collection of information consisting of numerical values or evidential facts, which can be utilised for analytical and inferential purposes. The term “data”, in the plural, comes from the Latin “datum”, meaning “things given or granted”. The term entered the English language in the late 1640s and, during the seventeenth and eighteenth centuries, was used for theoretical mathematical writing (Floridi, 2021; Rosenberg, n.d.). Data can be presented in both quantitative and qualitative forms, serving as a primary source of raw information about a specific object or phenomenon. This information is subsequently incorporated into analytical frameworks, ultimately contributing to the development and validation of theoretical constructs within the theory of knowledge (Wolski and Gomolińska, 2020).

This section provided a fundamental understanding of the concept of data; however, its relevance to the present discussion warrants further exploration. Threat actors often engage in unethical data acquisition from institutions for illicit purposes. A notable case exemplifying this behaviour involves a now-dissolved data science company (GOV.uk, n.d.) that unlawfully obtained data from the social media platform Facebook. Subsequently, this organisation harvested additional user data without obtaining user consent, not for monetary gain, but to advance specific agendas or objectives on behalf of their clients. This case highlights the ethical violations associated with unauthorised data collection and its potential implications.

It is generally assumed that a registered organisation would adhere to established ethical standards. However, the actions of this company, which manages user data, raise questions about the ethical considerations guiding its practices. Specifically, why did the company choose to circumvent ethical principles by utilising user data to advance its clients’ agenda? Furthermore, the legitimacy of the agenda they aimed to support warrants critical examination.

2. Origin of the Study

Christopher Wylie, a whistleblower who worked for the infamous Cambridge Analytica (CA), describes the company as not merely a data and analytics firm; instead, it functions as a mechanism explicitly engineered to disseminate disinformation. Wylie explicates that during his employment with the company, they exploited Facebook's platform to collect data from millions of users without their informed consent. This data was utilised to construct detailed profiles intended to influence users' opinions on specific topics and agendas (Cadwalladr and Graham-Harrison, 2018; The Guardian, 2018).

A critical issue concerns how a large social media organisation such as Facebook, with over three billion active monthly users and recognised as the first platform to reach one billion registered users (Dixon, 2025c), enabled third-party entities, including Cambridge Analytica, to harvest substantial volumes of user data for strategic purposes. Facebook's data holdings encompass a wide range of personal information, including names, ages, contact details, account credentials, profile photographs, follower networks, and behavioural data collected from both online and offline activities via cookies (Meta, 2018; Meta, 2025a).

It is widely argued that social media platforms such as Facebook should implement comprehensive strategies to protect user data from breaches by both threat actors and third-party applications. Notably, the data harvesting in this case was conducted by a UK-registered company rather than by external malicious actors. This distinction is significant, as it highlights the risk that threat actors could impersonate legitimate organisations to facilitate unauthorised data extraction for illicit objectives.

The ethical obligations of third-party entities, such as Cambridge Analytica, which harvested data from Facebook and subsequently utilised this information to promote agendas, prompt critical ethical questions about the deployment of data in the dissemination of misinformation. This paper endeavours to rigorously analyse the moral responsibilities and societal implications of data science companies that influence public opinion through Facebook campaigns, with a specific focus on relevant legislative frameworks in the United Kingdom and the societal repercussions arising from the scandal.

3. Background to the Study

Social media has become increasingly normative in contemporary society, significantly altering patterns of communication and interpersonal interaction. Users now leverage social platforms as primary sources for information, news, and entertainment, as platforms continuously develop innovative strategies to enhance user engagement and connectivity (Chinthala, 2023; Seng et al., 2021). Facebook is the world's largest social platform, boasting over three billion active users worldwide (Dixon, 2025c). Facebook platform gathers extensive data on its users, encompassing behavioural patterns, social network connections,

product utilisation, transactional activities within the platform, and analyses of user interactions within their networks for targeted advertising purposes (Meta, 2022).

Although Facebook's ongoing innovations and achievements demonstrate a proactive approach towards future technological advancements, the platform has encountered several ethical issues. Specifically, Facebook utilises third-party applications that, upon user interaction, access and potentially exploit users' social and personal data. Notably, the data science company, Cambridge Analytica, engaged in such practices by harvesting data not only from individuals who interacted with their application but also from the wider network of those users, raising significant concerns regarding data privacy and ethical responsibility (Cadwalladr and Graham-Harrison, 2018; Ur Rehman, 2019; Seng et al., 2021).

The data science company purportedly aggregated user data and utilised psychometric assessments to construct user profiles to predict behavioural patterns. This information was allegedly employed to influence political outcomes, particularly during the 2016 United States presidential election, wherein social activity data was leveraged to sway voter behaviour (HIIG, 2018; Hu, 2020; Schneble et al., 2018). This highlights the ethical shortcomings displayed by both Facebook and data science companies that claim to follow ethical standards. Such lapses have not only caused reputational damage to Facebook but also exposed significant flaws in data governance and accountability.

The data collection practices of the data science company highlight the absence of an ethical framework within major institutions that manage extensive user data. These organisations bear the responsibility of implementing robust safeguards to protect user information from malicious actors and unauthorised third-party applications, including the data science company in question. Additionally, organisations such as Facebook are urged to improve transparency regarding their data management practices. According to Benesch (2021), concerns exist that such corporations prioritise profit over their users' well-being, with their internal operations remaining largely opaque to external scrutiny.

This paper critically examines the ethical practices of the data science firm Cambridge Analytica, especially its data harvesting from Facebook. The company collected data to disseminate targeted advertisements, raising concerns about their authenticity and potential to spread misinformation. Furthermore, the paper discusses Facebook's apparent lack of ethical oversight, which facilitated the harvesting of user data and the platform's alleged support for CA's dissemination of misinformation. It also addresses the broader societal implications of data-driven campaigns and emphasises the necessity of establishing a comprehensive ethical framework to mitigate associated risks.

4. Problem Statement

The allegation concerning CA, a UK-registered data science company, harvesting data from Facebook's platform raises significant ethical questions, data privacy concerns, and

misinformation. The company purportedly extracted extensive quantities of personal data from Facebook users, subsequently utilising this information to develop targeted advertising campaigns. Notably, these campaigns appear to have lacked rigorous fact-checking by Facebook.

According to the BBC (2018a), an interview with Christopher Wylie reveals that CA, a data analytics company he was previously associated with, spread misinformation to unsuspecting users, particularly during the United States presidential election campaign. This was achieved through targeted advertising strategies that influenced user behaviour. By aggregating data from Facebook users, the organisation not only accessed information from applications utilised by unsuspecting individuals but also extracted data from the social networks associated with these users. This method could have expanded the campaign's reach, spreading misinformation further.

Moreover, if such campaigns achieve substantial engagement, Facebook's AI algorithms may amplify this content, potentially prioritising posts that could generate financial gains for the platform. This raises significant concerns regarding the manipulation of information dissemination and the platform's fiduciary responsibilities. According to O'Hagan (2018), Facebook has been accused of controlling the types of content displayed on its platform, prioritising content that enhances profitability and user engagement.

Several concerns have been articulated concerning the ethical implications of the company's data collection practices on social media, especially concerning data privacy and content integrity. The data science organisation, according to Wagner (2021), states that the data harvesting constituted a breach of privacy and therefore breached data protection legislation, allowing misinformation from targeted ads. In this context, Facebook users had their data unlawfully collected without consent and subsequently targeted with misinformation advertising.

This situation also amounts to negligence by Facebook, as legislative frameworks require institutions like Facebook to protect users' data. The platform's architecture allowed external developers to access user data, facilitating the data collection efforts of Cambridge Analytica and resulting in the compromise of information belonging to millions of users.

Facebook's failure to adhere to data privacy regulations was exploited by a data science company. Facebook had an obligation to adhere to legal frameworks such as the OECD (Organisation for Economic Co-operation and Development) Privacy Guidelines, which constitute the pioneering international framework for data protection, and the General Data Protection Regulation (GDPR) established within the European Union. Both set out the responsibilities of organisations that hold public data, such as Facebook, to safeguard users' data (González-Pizarro et al., 2022; Wagner, 2021; OECD, 2023). It can be argued that the data-harvesting practices of the data science company, along with Facebook's system allowing third-party access, demonstrate a lack of responsibility. This situation increases the risk of data breaches, jeopardises user privacy, and erodes trust in the platform.

This paper examines the ethical considerations inherent in the practices of data science companies and social media platforms, especially Facebook. The proliferation of misinformation highlights the urgent need to develop stronger methods to prevent ethical violations.

5. Aim

This paper aims to critically examine the ethical obligations and societal implications associated with data science companies, particularly Cambridge Analytica, and its influence on users through social media campaigns on platforms such as Facebook.

6. Objectives

1. To investigate how the data science company harvested data from Facebook to launch targeted campaigns aimed at influencing user behaviour towards a particular outcome.
2. To identify the issues and challenges related to ethical practices and professionalism within the data science industry, and to understand their influence on the field.
3. To analyse the impact of the campaigns conducted by the data science company on its target audiences to accomplish its strategic objectives.
4. To formulate a comprehensive theoretical and conceptual framework aimed at addressing the adverse effects of campaigns directed towards specific populations or issues.

7. Research Question/Hypotheses

1. Does a data science company harvest/gain audience data ethically?
2. What are the implications/impact of harvesting/gaining audience data unethically?

8. Significance of the Study

This study primarily investigates the ethical considerations pertinent to a data science company's collection of data from Facebook, a social media platform with over 3 billion users, without users' consent. It examines the implications of utilising this data to develop targeted marketing strategies tailored to individual social behaviours. The research aims to address critical issues concerning privacy, transparency, and the proliferation of misleading content, with reference to the CA's scandal (Townsend and Wallace, 2016).

This study will explore the ethical frameworks governing social media to evaluate current ethical standards. It will examine the social impact and public trust related to data-driven campaigns that influence user behaviour. Particular attention will be given to the ethical considerations surrounding the dissemination of both misinformation and accurate content. Additionally, the study will analyse relevant policies and legal frameworks, including best practices and potential mitigation strategies to address ethical challenges in digital content dissemination (Lynwood, 2025).

9. Scope of the Study

This study aims to explore the ethical considerations involving a data science company and its influence on users through social media campaigns on the Facebook platform. The investigation will scrutinise the company's unethical practices, including the unauthorised collection of user data from Facebook without consent, the dissemination of misinformation through targeted campaigns, and Facebook's role in enabling and potentially facilitating these activities. The analysis seeks to provide a comprehensive understanding of the ethical implications of these actions and the platform's involvement (Lauer, 2021).

The study will also review current academic literature, focusing on the theories and ethical debates concerning the ethical practices of both the data science company and Facebook, a social media platform.

The Cyber Security Body of Knowledge (CyBOK) serves as a comprehensive framework designed to facilitate education and training in the cybersecurity field (CyBOK, n.d.). This academic article examines the case of Cambridge Analytica, a data analytics company involved in data misuse, within the context of the CyBOK framework's emphasis on human, organisational, and regulatory factors. The discussion aims to critically analyse the ethical, legal, and societal implications of Cambridge Analytica's practices in relation to Facebook, emphasising human and organisational responsibilities in safeguarding cybersecurity.

This study does not endeavour to present a comprehensive forensic analysis of the scandal, primarily due to data limitations. Instead, it concentrates on the ethical, professional, transparency, and social implications, exploring how data science firms utilise targeted advertising to disseminate misinformation and the resulting consequences.

10. Rationale of the Study

The primary justification for these studies is rooted in examining the ethical considerations associated with data science corporations' practices of collecting users' data without explicit consent on social media platforms such as Facebook. Subsequently, this data is utilised to disseminate targeted campaigns. Facebook has faced significant criticism for inadequate data

protection measures, which constitute a foundational aspect of its operational model (HIIG, 2018; Lauer, 2021). Furthermore, allegations against both Facebook and data-harvesting entities highlight regulatory inaction and the absence of governmental oversight of Facebook's data management practices (Zinolabedini and Arora, 2019).

11. Structure of the Dissertation

The structure of this paper is divided into three distinct chapters. The initial chapter provides a comprehensive overview of the research scope, including the study's objectives, employed methodologies, aims, and underlying rationale. The second chapter presents a critical review of the existing literature, drawing on sources from both reputable news outlets and scholarly publications, thereby reflecting the interdisciplinary nature of the topic. The final chapter synthesises the collected data and insights, employing a qualitative research approach and a pertinent artefact to illustrate potential mitigation strategies.

12. Summary

Data has become an integral component of contemporary society, with prominent institutions such as Google, Amazon, and Facebook possessing extensive repositories of user information, including behavioural patterns, preferences, dislikes, social trends, and network activity. This paper will critically analyse the allegations made by the data science firm Cambridge Analytica regarding its purportedly unethical methods for extracting user data from Facebook and examine the extent to which Facebook, as a large organisation, permitted such practices. Cambridge Analytica utilised psychometric testing to develop detailed user profiles aimed at predicting behavioural tendencies, which were subsequently employed to target users with advertisements that were often misleading or deliberately misinformative. This issue raises significant ethical concerns, particularly regarding the proliferation of misinformation, which can have tangible impacts beyond the digital sphere, affecting individuals emotionally and physically. Furthermore, the paper will examine the allegations that Facebook failed to adequately safeguard user data, as its system allowed third-party applications to access personal information, thereby revealing apparent lapses in corporate responsibility. These issues have catalysed calls for the development and implementation of a more robust data protection framework to enhance user privacy rights.

Chapter 2:

13. Literature Review Introduction

This chapter critically reviews the literature on the ethical issues and societal impacts of data science companies, especially those targeting Facebook users with microtargeted ads. As digital tech and social media expand, organisations use advanced analytics and algorithms to influence user experiences and public discourse (Adeniran et al., 2024; Afsharian, 2025). The debate on CA and increased scrutiny emphasises the need to reassess corporate practices and responsibilities (Albright, 2018; Barrett, 2018).

It examines data-driven influence through targeted ads, focusing on ethics, privacy, governance, and political impacts. It analyses Facebook's data handling, considering localisation, privacy paradoxes, consent, regulation gaps, accessibility issues, and misinformation spread (Barth and de Jong, 2017; Benesch, 2021).

The chapter is organised into:-

1. Impact of digital tech on society and culture
2. Data privacy and ethics in the platform economy
3. Governance, regulation, and transparency challenges
4. Political implications and data manipulation.

14. Background: Technology, Society, and Data

14.1. Technology Proliferation and Social Transformation

The advent of technological advancements has profoundly transformed societal structures, with the proliferation of Internet of Things (IoT) devices serving as a key factor enabling the internet to connect to sensors and computers within the household local area network (LAN) (Krainyk et al., 2019). This evolution has not only changed how technology is utilised but also fundamentally reshaped societal interactions and engagement. According to Alsaleh (2024), the integration of technology within cultures is an important area of research, demonstrating its capacity to reshape cultural norms. This assertion is supported by Lusha (2023), who states that societal consumption of technology has fostered a normative sociality, thereby altering traditional models of information acquisition. An example of this is during the COVID-19 pandemic, when many people worked from home (WFH), utilising tools such as Zoom, MS Teams, Slack, and Google Chat, whilst spending more time on social media platforms to stay in touch with colleagues, friends, and family members (Lal et al., 2021).

14.2. Data Generation and the Use of Social Media

The use of technology, particularly on social media platforms such as Facebook, has led to a substantial increase in data collection. Facebook, with its user base exceeding three billion active users, facilitates the real-time sharing of extensive personal information. According to Stieglitz et al. (2018), social media has become an integral component of contemporary daily life. Its widespread popularity can be attributed to its cost-effectiveness, essentially free access, and high accessibility, which collectively drive a significant increase in data acquisition. This trend is further corroborated by GWI (2025), which, in its 2025 report, states that the popularity of social media continues to grow, with an increasing emphasis on video-based content as the primary medium for information dissemination.

14.3. Demographic and Accessibility

Facebook's global user base encompasses individuals aged 18 and over. The predominant age group is 25 to 34 years old. Within this demographic, male users make up approximately 18.5%, while female users account for 12.7%. A significant majority of users, 98.5%, access the platform via mobile devices. The average daily engagement is about 31 minutes per day (Dixon, 2024; Dixon, 2025a; Sheikh, 2025). These data highlight the platform's high accessibility, especially among male users, due to mobile optimisation and worldwide reach. In 2020, Asia led with over 823 million active Facebook users, including 375 million from India. Latin America and Europe followed, with over 414 million and 395 million users, respectively. The Middle East and Oceania/Australia had fewer users, with 93 million and 23 million, respectively (Dixon, 2025b; Kumar, 2025).

The data suggest Facebook has a worldwide reach with a predominantly male user base, reflecting the 2020 population of approximately 3.95 billion males and 3.9 billion females (Galan, 2025).

A company with an extensive user base and international presence must prioritise maintaining active engagement with its platform. Facebook's algorithms rely on data collected from users, encompassing their preferences, interests, and online behaviours. This data enables Facebook to deliver highly targeted advertisements tailored to individual user profiles, thereby enhancing advertising efficacy through behavioural targeting (Hitlin and Rainie, 2011, pp.1–23).

14.4. Algorithmic Content Curation and Ethical Issues

Questions emerge about Facebook content dissemination, especially considering its algorithms and AI models. A key concern is verifying the accuracy of online information. Lauer (2021) criticises Facebook for spreading misinformation. Facebook's Chief AI Scientist, Yann LeCun, claims that the company's AI filters help prevent disinformation.

Nonetheless, Lauer (2021) highlights a fundamental issue rooted in Facebook's profit-driven model that may encourage the spread of viral misinformation through AI. Zollo and Quattrociocchi (2018) add that user engagement also spreads false information, as users interact with it more, thereby boosting virality. This is worsened by engagement-optimising algorithms, raising ethical issues. A significant challenge is researchers' limited access to Facebook's algorithms, since Facebook keeps its AI details private (Lauer, 2021). While companies have the right to protect their innovations, Lauer argues that Facebook's management of data from over a billion users merits scholarly review to understand its societal impact.

Benesch (2021) emphasises that employees must sign nondisclosure agreements. Whistleblower Frances Haugen claimed Facebook prioritises profits over user wellbeing, allowing false, misleading, or toxic content, and censors millions of posts daily. The platform's news highlighting decisions are made by Facebook's internal team and AI algorithms, which are opaque, raising ethical concerns. Such content can significantly impact social and psychological well-being, emphasising the need to understand these processes.

14.5. The Role and Ethical Obligation of Data Science Companies

Given the ongoing ethical concerns surrounding Facebook's use of its AI models to preferentially promote certain content while restricting others, to what extent could this knowledge potentially be exploited, and did a data science company use this to their advantage?

The primary role of a data science company is to perform comprehensive analyses of complex datasets to support the organisation's strategic goals. This involves using advanced analytical methods to foster innovation and improve operational efficiencies, thus supporting data-driven decision-making and boosting organisational effectiveness (Adeniran et al., 2024; Afsharian, 2025).

As corporations embrace digital technologies, data availability across sectors increases. However, managing sensitive information demands strict ethical standards and best practices in data handling. According to Adeniran et al. (2024), there is a growing recognition within academic and professional circles of the importance of ethics in data science. As data collection broadens, it is crucial to develop a comprehensive ethical framework to ensure compliance, safeguard privacy, enhance transparency, and prevent biases, thereby reducing unethical practices (Dhiman, 2023).

14.6. Theoretical and Conceptual Foundations

Institutional, professional, and legal frameworks predominantly govern the ethical management of data within data science organisations and digital platforms. The British

Psychological Society (2021), alongside comprehensive guidelines on internet and social media ethics from universities such as LSE and the University of Glasgow (Lynwood, 2025; University of Glasgow, 2025), emphasises the primacy of user consent, transparency, and privacy. These principles are reinforced through institutional oversight mechanisms, such as Institutional Review Boards (Grady, 2015), and the formulation of robust ethical standards for the handling of personal and sensitive data (Markham and Buchanan, 2012).

Furthermore, ethical decision-making in data science is guided by prominent theoretical frameworks, such as Utilitarianism. This theory focuses on maximising overall benefit for the majority; however, its application may potentially compromise individual privacy rights (Deuker, 2010). Deontological ethics, which stress duties and rights, underscores the significance of respecting users' autonomy and ensuring informed consent, which must be upheld (British Psychological Society, 2021). Principlism outlines core ethical principles, respect for individuals, beneficence, non-maleficence, and justice, thereby offering a comprehensive framework for evaluating data practices (Lindridge, 2017). Privacy Calculus assesses the balance between potential risks of data disclosure and anticipated benefits, profoundly influencing user behaviour, decision-making, and organisational data strategies (Barth and de Jong, 2017; Choi et al., 2018).

Platforms such as Facebook have established a new form of economic and social organisation, termed “surveillance capitalism,” where data is regarded as a fictitious commodity, generated and exchanged with little oversight or due diligence (Manokha, 2025). The consequences of such platform capitalism practices include prioritising user engagement and data monetisation, often at the expense of user privacy (Lauer, 2021; Stieglitz et al., 2018).

A further layer of regulation is provided by legal instruments such as the General Data Protection Regulation (GDPR) in the EU, which formalises users' rights regarding their personal data. Additionally, the UK and other jurisdictions have imposed hefty fines on companies like Facebook and those involved in the Cambridge Analytica scandal (Davies and Rushe, 2019; Hern, 2019; ICO, 2023).

Although a regulatory framework exists, transparency issues remain concerning algorithms and proprietary AI, raising concerns about bias, misinformation, and manipulation of user behaviour (Benesch, 2021; Lauer, 2021). Moreover, the practical application of the GDPR's focus on explainability and user control remains insufficient.

These ethical theories and legal frameworks form a fundamental basis for assessing the practices of digital platforms and data science companies. A clear gap persists between these normative guidelines and their real-world implementation, highlighting the ongoing need for vigilance and accountability. Therefore, enhancing transparency and establishing enforceable policies concerning algorithmic content curation and data management are imperative.

15. Who is Cambridge Analytica

Cambridge Analytica, a British analytics firm founded in 2013 with Alexander Nix as CEO, is a subsidiary of the SCL Group, formerly known as Strategic Communication Laboratories. It concentrated on U.S. elections and received approximately \$15 million from Republican supporter Robert Mercer. The SCL Group has operated for over 25 years (Cadwalladr and Graham-Harrison, 2018; Fernando, 2021; Ingram, 2018; Siegelman, 2018).

The company's websites, sclgroup.cc and cambridgeanalytica.org, are currently non-operational, making verification difficult. Archived records show SCL claims over 25 years of activity, specialising in behavioural interventions and collaborating with political entities worldwide, reportedly endorsed by agencies such as the Ministry of Defence, the US State Department, and NATO (Web.archive.org, 2016).

Chang (2018) describes the SCL group as a shell company, a point also made by Kenton (2019). A shell company is an entity with no active operations, assets, or employees. While not illegal, shell companies are often used to conceal ownership. They are sometimes utilised legitimately, such as by small startups raising capital (Kati, 2022).

The SCL Group has 18 subsidiaries in the UK and the US, complicating understanding of its structure, funding, and decision-makers. It also faces challenges in deciphering operational methods and data-sharing mechanisms (Siegelman, 2018). The group reports collaborations with government and military on research and counter-narcotics projects (Cadwalladr and Graham-Harrison, 2018). This secrecy suggests it handles sensitive information or engages in ethically questionable practices, possibly explaining CA's use of a shell company without direct ties.

16. The Cambridge Analytica Case: Data Harvesting and Exploitation

CA's scandal exemplifies the complex interplay between data science, social media, and political strategy in the contemporary digital landscape, where the extensive accessibility of data facilitates targeted outreach to specific demographics and enables personation, surpassing the capabilities of traditional media such as television advertising (Akbar et al., 2025; Filmology, 2025). The controversy surrounding CA's large-scale data harvesting and the exploitation of users' Facebook data has raised significant privacy concerns, questions about the integrity of online platforms, and the ethical implications of algorithmically driven advertising.

The data was collected through Aleksandr Kogan's Facebook application survey, titled "This Is Your Digital Life." This application not only collected data from survey respondents but also aggregated a variety of information from their social networks, including data on friends, likes, locations, and, in some instances, private messages (Barrett, 2018; Hartmans, 2018; Cadwalladr and Graham-Harrison, 2018; Hern and Cadwalladr, 2018). At that time,

Facebook's API system permitted third-party applications to access data from any user who provided consent, thereby significantly broadening the scope of data acquisition with minimal or no explicit consent from the users involved (Albright, 2018).

Once the application collected the data, it was repurposed for psychographic profiling (Bakir, 2020; Hu, 2020; Prichard, 2021). Utilising the OCEAN model of the “Big Five” personality traits, CA was able to infer users’ psychological predispositions and vulnerabilities. These profiles were then linked to microtargeted ads or messages crafted to elicit specific emotional and behavioural responses (Rosenberg et al., 2018; Sutton, 2025; The Guardian, 2018). Such practices are commonly associated with “psychological operations,” which integrate data science, behavioural economics, and strategic political campaign methodologies.

The magnitude and repercussions of the CA scandal are estimated to have impacted approximately 87 million Facebook users worldwide, including slightly over 1 million residents within the UK (Hartmans, 2018; The Guardian, 2018). The influence of CA extended beyond the US presidential election, with allegations suggesting that the organisation also affected the Brexit referendum and various elections across Africa, thereby shaping political discourse through targeted dissemination of misinformation (Dowling, 2022; Hunt and Messinger, 2018; Netflix, 2019). The scale of CA's operations underscores the capacity of such entities to deliberately manipulate political outcomes via data-driven psychological campaigns, raising profound ethical concerns and highlighting issues related to the misuse of technological resources and privacy rights.

17. Cambridge Analytica scandal overview

The controversy surrounding the data science company Cambridge Analytica originated with Aleksandr Kogan, an academic at the University of Cambridge. Initially, Alexander Nix, then the CEO of CA, contacted Julian Assange, the founder of WikiLeaks, regarding the hacked emails from the Democratic National Committee. Nix purportedly possessed access to emails pertinent to Hillary Clinton; however, Assange declined Nix's request (Ballhaus, 2017; Chang, 2018).

Alexander Kogan developed “This is Your Digital Life,” a quiz-based application on Facebook that used Amazon Mechanical Turk and Qualtrics. Participants were informed that their data would be used for academic research; approximately 270,000 Facebook users completed the quiz. However, the app also gathered data from their social networks, leading to unauthorised access to up to 87 million Facebook profiles, including individuals who did not participate or give consent, raising ethical concerns (Chang, 2018; Hunt and Messinger, 2018; Isaak and Hanna, 2018; Wong et al., 2018).

Kogan shared data with SCL Election, which was introduced to him in 2014 by a Cambridge PhD Psychology student. SCL Election, a subsidiary of the shell company conglomerate SCL

Group linked to CA, was involved in data sharing under psychometric research, with Kogan consenting mainly for financial incentives to survey participants.

Kogan's data-sharing with a third-party allegedly violated Facebook's terms of service. Initially, he claimed the data was for academic use, but later admitted it was transferred to GSR, a company he co-founded with Joseph Chancellor, a Cambridge researcher and Facebook employee. The relationship between Kogan and Facebook is unclear, but whistleblower Wylie indicated Facebook knew about the extensive data extraction by Kogan's app (Hern and Cadwalladr, 2018; Lewis and Wong, 2018; Wong and Lewis, 2018).

CA used this data to harvest information from approximately 87 million Facebook users, including over one million in Britain, without their consent, raising ethical and legal concerns (Chang, 2018; Hunt and Messinger, 2018; Kenber, 2018; Wong et al., 2018).

Utilising the newly acquired data, CA built psychographic profiles of Facebook users, assessing Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN). Isaak and Hanna (2018) note that the GSR needed to keep data to develop profiles with over 5,000 points. These profiles targeted ads to influence user behaviour, especially to sway voters in the 2016 US presidential election (Cadwalladr, 2019; Chang, 2018; Hunt and Messinger, 2018; Isaak and Hanna, 2018).

CA's psychographic profiling and targeted ads helped Donald J. Trump win the 45th U.S. President, defeating Hillary Clinton with 304 to 227 electoral votes (History.com Editors, 2018; The New York Times, 2017). After the 2016 US election, CA garnered media attention when whistleblower Christopher Wylie, a co-founder of the organisation, alleged they engaged in unlawful data harvesting from Facebook for psychographic profiling. This profiling reportedly aided targeted political adverts, influencing the election in Trump's favour. Wylie described the organisation as a "propaganda machine" spreading misinformation to sway public opinion (Cadwalladr, 2019; Chang, 2018; Hunt and Messinger, 2018; The Guardian, 2018).

Romano (2018) states that CA's success in harvesting Facebook data resulted from a vulnerability in Facebook's API. Third-party access enabled developers to gather data beyond user interactions, bypassing restrictions on social networks. Facebook was legally responsible for this breach, leading to a £500,000 (approximately \$663,000) fine by the ICO for failing to protect user data. This incident caused a \$119 billion loss in value and a decline in user engagement (Cadwalladr, 2019).

18. Facebook and its third-party application

According to Hartmans (2018), the specific types of data CA obtained from Facebook remain unclear due to the company's lack of transparency. Typically, users who engage with third-

party apps share information such as their location, platform activity (likes, shares), photos, and interests, all of which are accessible to the application developers.

Under the previous Facebook API framework, specifically the Graph API v1.0, third-party applications were granted access to user data, raising significant privacy and ethical data management concerns. This API model included an “extended permission” feature that not only permitted third-party applications to access primary user information but also allowed these applications to operate covertly in the background. Such applications could collect data about a user's network of friends without obtaining explicit user consent or providing any notification. Furthermore, the Facebook API v1.0 facilitated requests to access users' private message inboxes, also known as Facebook Direct Messages (DM), through the “read-mailbox” API endpoint, thereby exacerbating potential privacy infringements (Albright, 2018; Hartmans, 2018; Isaak and Hanna, 2018; Wong and Lewis, 2018)

According to Barrett (2018), Facebook notified users affected by the CA data scandal through a targeted message, including those who engaged with the quiz “This is Your Digital Life” or had data collected. The notice titled “Protecting Your Information” listed compromised data types such as public profile information, page likes, birth date, and location. Tsormpatzoudi et al. (2018) added that data also included details like user profiles ('about me', 'hometown', 'interests', 'relationship details', 'religion', 'politics', 'status', 'website', 'work history'), activity logs (actions, activities, birthday check-ins, history, events, games activity, groups, notes, online presence, photo/video tags, photos, questions, subscriptions), and social connections (friends, relationships). Facebook stated that the implicated app was terminated and advised users to review and update their privacy settings in “Apps and Websites” to remove unauthorised access. Notably, this setting was hidden from the main “Privacy” options until the scandal emerged.

According to Albright (2018), Facebook was fully aware of the extensive user data accessible through its initial Graph API v1.0. Despite this awareness, the company opted not to deactivate the system, as it served as a crucial channel for business partners and generated significant revenue. This decision suggests that Facebook demonstrated limited regard for user data protection. This is also supported by Barrett (2018), who states that Facebook's internal employees had identified ethical concerns within the platform but prioritised profit over safeguarding user information. Furthermore, this zero-day vulnerability could have been significantly more damaging had malicious threat actors exploited it, potentially surpassing the scale of previous breaches, such as the GSR and CA scandals.

Since the CA scandal, Facebook has updated its API from version 1.0 to 2.0. This major update imposes more stringent restrictions on applications accessing users' social networks, requiring explicit user authorisation for data access. Additionally, it enhances permissions granted to third-party applications and mandates a formal verification process for access to advanced features (Meta, 2025a). The implementation of this upgrade was a strategic response to the data breach, aimed at restoring trust within the platform's user community with a “People First” methodology (Constine, 2015).

19. Current studies on influence and ethics in data science

The internet has increasingly become a fundamental platform for data collection and research, encompassing a broad range of topics, including political orientation and user behaviour. Tools such as Amazon Mechanical Turk (MTurk), and survey platforms such as SurveyMonkey, Typeform, and Google Forms are essential for basic data gathering from participants. Furthermore, social media platforms serve as vast repositories of user-generated information, and users often cannot distinguish whether the content in their newsfeeds is factual, misleading, or designed to influence their opinions through targeted advertising. However, this rise in data collection raises significant ethical and technical concerns regarding data protection, user consent, and transparency. Addressing these issues requires adherence to the CIA triad framework for data management, which includes Confidentiality, Integrity, and Availability, to promote ethically sound and secure data practices (Amazon Mechanical Turk, 2018; Dhiman, 2023; Cawthra et al., 2020; Hinds et al., 2020; Schneble et al., 2018; Townsend and Wallace, 2016; Zapier, 2019).

Data protection has grown ever more crucial, not only for individuals whose data is being analysed but also for the researchers and institutions performing these analyses. Organisations with large datasets are required to comply with the General Data Protection Regulation (GDPR); however, there is a noticeable absence of a comprehensive framework delineating ethical guidelines for researchers analysing data sourced from social media platforms (Hinds et al., 2020; Townsend and Wallace, 2016; Tarran, 2018). According to Schneble et al. (2018), the data shared with Cambridge Analytica was anonymised and aggregated; however, Tsormpatzoudi et al. (2018), in their study, presented evidence suggesting that the data also encompassed personally identifiable information such as ‘about me’ sections, status updates, birthdays, and other personal details. When users register for online services, such as those provided by platforms like Facebook, they often disclose multiple pieces of personal information.

Frequent changes to settings and privacy policies complicate users' ability to fully understand and manage their online privacy (Hinds et al., 2020). This indicates a lack of transparency from Facebook and CA, which hindered researchers from conducting a thorough investigation. According to Tarran (2018), data scientists should adhere to an ethical oath when handling data, analogous to the Hippocratic Oath taken by physicians. This practice aims to ensure that data scientists comprehend and uphold ethical standards in their professional conduct.

When compared with traditional research disciplines such as medicine, ethical evaluation is an integral part of the research process, ensuring adherence to established best practices. In the field of data science, however, obtaining informed consent from participants presents notable challenges, particularly when data are derived from secondary sources. Such data may originate from a diverse array of sources, ranging from fully anonymised datasets to those containing identifiable information. Securing consent from every individual involved

requires direct contact, which becomes increasingly unfeasible given the large number of participants in data science projects (Schneble et al., 2018; Townsend and Wallace, 2016). An ongoing scholarly debate concerns whether data sourced from social media platforms should be classified as public or private information. This discussion primarily stems from the terms of service users agree to at the point of registration, which often delineate the conditions under which their data may be accessed by third parties (Townsend and Wallace, 2016). This issue alone engenders significant ethical considerations, given that data available on social platforms is publicly accessible and thus accessible to researchers and third-party institutions. However, researchers must remain cognizant of the ethical implications involved in accessing such data. According to the British Psychological Society (2021), researchers are expected to maintain a high standard of professionalism and adhere to ethical principles consistent with the Code of Human Research Ethics.

An additional concern for researchers collecting data from social media platforms concerns the degree of anonymisation of data about individual users and how they determine whether those users are minors. Social media platforms tend to retain data over long periods, and such data remains searchable. Users share significant amounts of information, some of which may be considered private. When shared with a select few within their social network, it remains in the social media database (Townsend and Wallace, 2016). The potential risks associated with this practice include system vulnerabilities that may arise if platforms do not regularly update or securely migrate their systems. These vulnerabilities can be exploited by threat actors to exploit zero-day exploits, potentially exposing large volumes of unanonymised user data that remain susceptible to breaches.

Schneble et al. (2018) researched ethical practices in ten universities across the US, UK, and Switzerland. Most lack comprehensive data science frameworks, reducing their ability to evaluate social media and Internet data ethically. They call for developing such frameworks, emphasising critical reasoning as per the Association of Internet Researchers (Markham and Buchanan, 2012). However, understanding legal and technical issues is necessary, which could create liabilities rather than safeguard ethics. Access to the data is under review, so further analysis is unavailable.

20. Ethical Implications of Data Misuse

The scandal involving CA has significantly raised public awareness of privacy issues on social media platforms like Facebook. Empirical research has examined differences between users' perceived risks and the actual threats they face. For instance, users frequently express concerns about privacy breaches, which can lead to incomplete disclosure of personal information (Barth and de Jong, 2017; Deuker, 2010). However, their online behaviour does not consistently align with these concerns (Barth and de Jong, 2017; Choi et al., 2018). Furthermore, investigations following the CA scandal have highlighted that many users remain unaware of how accessible their personal data is to third parties. Even those who are aware often feel they lack control over protecting their data and tend to underestimate the

risks posed by advanced predictive analytics and psychometric profiling technologies (Alsaleh, 2024; Hinds et al., 2020).

Universities and research institutions are increasingly subject to scrutiny to establish comprehensive frameworks for ethical data science practices, aligning with guidelines posited by the British Psychological Society (2021), the Association of Internet Researchers (Markham and Buchanan, 2012), and local Institutional Review Boards (IRB) committees (Grady, 2015). Despite notable advancements, significant discrepancies persist between established academic ethical codes and actual research practices. According to Schneble et al. (2018), many research institutions lack systematic policies for assessing the ethical use of data obtained from social media sources. These challenges are exacerbated within the commercial sector, where organisations often prioritise proprietary interests and profitability over adherence to ethical standards (Epstein and Medzini, 2022; Lynwood, 2025). The proliferation of data-sharing mechanisms between academia and corporate entities, enabled by APIs and third-party platforms, raises complex issues related to research transparency, user protection, and institutional accountability.

Social media platforms, particularly Facebook, play a pivotal role in shaping perceptions of trust, transparency, and ethical standards within the digital ecosystem. Following revelations about CA practices, Facebook faced widespread international criticism and substantial fines for failing to safeguard user data (Barrett, 2018; Cadwalladr, 2019; Davies and Rushe, 2019; Hartmans, 2018). Investigative reports indicated that ambiguous API access and policy frameworks enabled large-scale data harvesting without user consent or raising user awareness (Albright, 2018). Although Facebook subsequently implemented more restrictive API access policies, considerable scepticism persists regarding the sufficiency of these platform-driven reforms in addressing underlying data privacy concerns (Benesch, 2021).

Data science firms face significant challenges due to evolving business models and scandals that have precipitated bankruptcies, eroded trust, and caused reputational damage, as exemplified by the case of CA (Kenber, 2018). A critical question that arises is: who bears responsibility for safeguarding user interests? Is it the platform itself (such as Facebook), the data science organisations, or the regulatory bodies? Addressing this question is essential for enhancing compliance, promoting ethical standards, increasing transparency, and encouraging active engagement with issues related to user rights and privacy (British Psychological Society, 2021; Tech Policy Press, 2024).

21. Political and Societal Impacts

The Cambridge Analytica scandal exemplifies a novel form of digital intervention in electoral processes. It underscores the potential utilisation of psychometric microtargeting techniques to disseminate misinformation and influence democratic decision-making. Documented evidence indicates that CA employed Facebook user data to develop highly targeted political advertisements during the 2016 US Presidential Election, with the explicit aim of

manipulating voter behaviour to favour specific electoral outcomes (Cadwalladr and Graham-Harrison, 2018; Hunt and Messinger, 2018; The New York Times, 2017). A comparable phenomenon has emerged in various regions globally, indicating that influence targeting in Australia, several African nations, as well as in Mexico, Brazil, India, and Malaysia, has involved digital microtargeting campaigns employed for political and social manipulation (BBC News, 2018b; Netflix, 2019; Nyabola, 2019; SBS News, 2018). This indicates that microtargeting and data-driven misinformation are not merely isolated phenomena restricted to specific geographical or cultural contexts but rather represent an emerging global trend.

The scandal involving Cambridge Analytica has catalysed a robust debate within academic circles about the ethical implications of digital technology, particularly regarding issues of responsibility and data governance. In the United States, congressional hearings have scrutinised the roles of key stakeholders from Facebook and Cambridge Analytica, while parliamentary inquiries in the UK have similarly interrogated these entities. These proceedings have highlighted critical questions concerning data ethics in the context of political processes (BBC News, 2018a; Confessore, 2018; Epstein and Medzini, 2022). This has prompted numerous inquiries concerning the impact not only of regulatory initiatives but also of the transparency practices employed by technology corporations in elucidating their internal policies. Specifically, current research suggests that these companies may face restrictions in scrutinising Facebook's AI model (Benesch, 2021; Lauer, 2021).

Such controversies pose a fundamental threat to democratic legitimacy, according to Dowling (2022). The use of data-driven psychological campaigns can destabilise the trust that underpins electoral processes and civil society. Responses from officials and stakeholders include a range of measures, such as regulatory interventions to address privacy breaches, alongside public initiatives to enhance digital literacy, raise awareness, and reinforce data protection standards (Federal Trade Commission, 2019). Many researchers argue that such responses remain fragmented, as academics and policy discussions increasingly advocate for holistic approaches that balance technological innovation with ethical considerations and the ongoing relationship between technology, commercial interests, and democratic accountability (British Psychological Society, 2021; Heawood, 2018).

22. Responses, Governance, and Future Directions

Following CA's scandal, a substantial response was mobilised across various social media platforms, notably Facebook. A historic legal action was initiated against Facebook, resulting in a \$5 billion fine levied by the U.S. Federal Trade Commission (FTC) and a £500,000 fine imposed by the Information Commissioner's Office (ICO) in the UK (Davies and Rushe, 2019; Federal Trade Commission, 2019; Hern, 2019). The sanctions imposed by these regulatory bodies underscore the social media conglomerate's deficiencies in safeguarding user data, marking a pivotal juncture in holding such platforms accountable. Consequently, in response to Facebook's inadequate data protection measures, the company expedited modifications to its API, including restricting access and eliminating friend-data sharing for

third-party applications, and implemented enhanced internal governance procedures for data permissions (Albright, 2018; Constine, 2015; Meta, 2025a). However, it can be argued that these interventions were predominantly reactive, serving to mitigate public perceptions of reputational damage; ongoing debates persist regarding the transparency of Facebook's continuous data practices (Benesch, 2021; Hartmans, 2018; Lauer, 2021).

Along with developments in platform technology, significant reforms to regulatory duties have been introduced to address challenges in global data governance. The General Data Protection Regulation (GDPR) within the European Union has established foundational principles such as enhanced user rights, data portability, and stringent consent requirements, thereby setting a precedent for best practices worldwide (ICO, 2023). The UK's Information Commissioner's Office (ICO) has provided comprehensive guidance tailored for businesses, academic institutions, and general researchers. Concurrently, online safety and privacy legislation continues to evolve to address emerging technological and societal needs (ICO, 2023; Legislation.Gov.UK, 2023; Marotta and Madnick, 2025). Despite these advancements, regulatory frameworks frequently lag technological innovation, and enforcement mechanisms vary considerably across jurisdictions. Some scholars argue that attempts to harmonise a unified global framework within a fragmented policy landscape may inadvertently create loopholes and undermine the consistency of user protection (Marotta and Madnick, 2025).

Both technical and social transformations necessitate that thought leaders advocate for adaptable, context-sensitive ethical frameworks, particularly for researchers engaging with large datasets (British Psychological Society, 2021; Markham and Buchanan, 2012; University of Glasgow, 2025). It is recommended that an explicit informed consent process be incorporated, alongside third-party data audits, rigorous anonymisation procedures, and transparency-by-design principles in both platform development and research methodologies (Lynwood, 2025). Furthermore, researchers and regulatory authorities have emphasised the importance of independent certification of platform practices, empowering users with autonomy over their privacy settings, and establishing precise mechanisms for redress in instances of misuse (EDHEC Online, 2025; Lynwood, 2025; Markham and Buchanan, 2012).

Despite notable advances, several enduring and significant gaps remain. These core challenges encompass issues related to legislation and enforcement within an increasingly fragmented regulatory landscape, characterised by substantial variability across jurisdictions. A significant gap in current research is the absence of comprehensive studies evaluating how effective existing regulatory frameworks are at protecting user data across international borders. Global data flows continue to bypass local protections in certain instances, facilitating cross-border data transmission (Marotta and Madnick, 2025; OECD, 2023). The ambiguity surrounding responsible entities arises from the absence of clearly delineated boundaries among platforms, data science organisations, and state authorities, resulting in a diffusion of responsibility and the emergence of a regulatory grey zone (Leonelli, 2016). Furthermore, platforms continue to maintain disparate data policies with minimal third-party oversight, and external academics and researchers frequently encounter restrictions on API access. These limitations inhibit independent auditing, large-scale research endeavours, and

transparency regarding institutional models (Benesch, 2021; Stieglitz et al., 2018). The phenomena of the “privacy paradox” and “privacy fatigue” exemplify users' perceived lack of control over their personal data, despite ongoing technical and regulatory reforms (Barth and de Jong, 2017; Choi et al., 2018).

Ongoing scholarly inquiry and policy discourse should prioritise the formulation of a balanced global framework that fosters collaboration among public entities and stakeholders across diverse regions and jurisdictions. This necessitates the development of innovative technical and legal instruments to ensure independent oversight. Attaining substantive, responsible data governance in the increasingly expansive digital environment requires a cross-disciplinary, multi-sector approach.

23. Cross-Sectional Synthesis

The consensus across scholarly, policy, and industrial domains emphasises that data privacy remains a vital and unresolved issue in the digital age. Empirical studies highlight the prevalence of “privacy fatigue,” a phenomenon in which users feel helpless due to complex privacy settings and recurring data breaches (Choi et al., 2018). This notion aligns with the “privacy paradox,” a phenomenon in which individuals express concerns about their privacy yet often fail to take protective measures to safeguard their online information (Barth and de Jong, 2017; Epstein and Medzini, 2022). Furthermore, it is widely acknowledged that users generally possess insufficient knowledge and awareness concerning the mechanisms of data collection, algorithmic processing, and monetisation practices involving their personal information (Deuker, 2010; Hinds et al., 2020).

The literature reveals notable divergences regarding both the efficacy and associated risks of psychographic profiling and microtargeting campaigns. Certain industry stakeholders and segments of the advertising sector perceive microtargeting as a natural extension of traditional marketing practices, emphasising its efficacy and personalised approach (Hewood, 2018). In contrast, recent empirical studies, particularly in the aftermath of the CA's scandal, highlight significant dangers, including psychological manipulation, the dissemination of propaganda, and the proliferation of misinformation (Dowling, 2022; Prichard, 2021). Furthermore, a discernible discrepancy exists between users' perceptions that “It wouldn't happen to me” and the concerns expressed by researchers, regulators, and activists regarding data misuse and algorithmic targeting. Notably, companies such as Facebook often withhold their models from external research, further exacerbating transparency issues (Benesch, 2021; Hinds et al., 2020).

The ongoing tension among the self-regulatory aspirations of social media corporations, emerging governmental policies, and the ethical considerations prioritised by academia highlights the persistent challenges inherent in this domain. Empirical studies have indicated that institutions such as Facebook and Google publicly espouse commitments to safeguarding user privacy; however, these declarations are often contrasted with their continued reliance

on core business models rooted in data monetisation and the deployment of opaque algorithms (Epstein and Medzini, 2022; Lauer, 2021).

Legislative initiatives, such as the UK's Online Safety Act 2023 or GDPR, constitute noteworthy advances; yet regulatory frameworks frequently lag technological developments and struggle to comprehensively address the intricacies of algorithmic targeting and cross-platform data exchanges (Dhiman, 2023; Legislation.Gov.UK, 2023).

Moreover, the rapid evolution of technology exacerbates regulatory oversight challenges, as government agencies often face bureaucratic inertia and must engage in extensive consultation before enacting legislation. Ethical oversight mechanisms within academia, such as Institutional Review Boards (IRBs), are predominantly designed for biomedical research and are inadequately equipped to address the unique ethical dilemmas presented by contemporary digital data practices. Many universities lack comprehensive ethical guidelines tailored explicitly to large-scale online datasets, frequently relying on traditional frameworks established for medical research (Grady, 2015; Schneble et al., 2018). According to Schneble et al. (2018), IRBs are ill-equipped to navigate the complexities and ambiguities inherent in research involving data collected from social and online platforms, where the delineation between public and private information is often ambiguous.

This study is situated at the crossroads of ongoing scholarly debates concerning digital psychometric profiling and its ethical, regulatory, and societal implications. By integrating technical analyses of psychometric techniques with a critical synthesis of ethical frameworks, regulatory gaps, and their real-world consequences, this research endeavour aims to address gaps in the existing literature. Notably, these gaps are highlighted through the exploration of the "privacy paradox" and fatigue phenomena, which significantly influence user behaviour. The use of psychographic profiling, supplemented by empirical data, helps distinguish between legitimate personalisation efforts and the spread of harmful misinformation. Furthermore, the study emphasises the need to develop comprehensive ethical guidelines and adaptive regulatory strategies that evolve alongside technological advancements to ensure transparency and accountability in digital data practices. This chapter advocates for an interdisciplinary approach that includes a reassessment of ethical standards and encourages government bodies to adopt a proactive stance in response to cultural shifts within industry practices and user awareness. It argues that fostering a nuanced understanding of privacy, agency, and ethical stewardship among users is essential for maximising the societal benefits of data science, while concurrently reducing democratic and psychological risks associated with digital technologies (Adeniran et al., 2024; Bakir, 2020; Markham and Buchanan, 2012).

24. Conclusion

The increasing prominence of social media and digital platforms has heightened awareness and spurred scholarly investigation in this domain. Contemporary researchers and academics are actively monitoring these developments, emphasising that while such platforms confer

significant socio-economic benefits and facilitate unprecedented levels of connectivity, they concurrently give rise to pressing ethical and regulatory challenges concerning users' privacy, informed consent, and data management (British Psychological Society, 2021; Lauer, 2021; Markham and Buchanan, 2012). Regulatory frameworks such as the EU's GDPR and high-profile investigations, notably into the Cambridge Analytica data harvesting scandal, have initiated a range of policy responses. However, these initiatives remain somewhat fragmented both within the EU and globally (Davies and Rushe, 2019; Hern, 2019; ICO, 2023). Various theoretical paradigms, including Utilitarianism, Deontological ethics, Principlism, and privacy calculus, shape governance approaches and underscore the inherent tension between technological innovation and the imperative to protect individual rights (Barth and de Jong, 2017; Deuker, 2010).

Despite notable advancements, the extant literature persistently highlights enduring deficiencies in the articulation of ethical guidelines. These inconsistencies are pervasive across the private, commercial, and research sectors, with legislative frameworks often lagging behind the rapid evolution of technological innovation. Consequently, user data remains susceptible to exploitation, manipulation, and extensive misuse (Benesch, 2021; Schneble et al., 2018). Platforms such as Facebook and third-party data brokers continue to operate within a regulatory "grey area," employing opaque algorithms and permissive data-sharing APIs (Albright, 2018; Lynwood, 2025). Users of these technologies display phenomena such as "privacy fatigue" and the "privacy paradox," expressing concerns about data security while remaining uncertain about effective protective measures (Barth and de Jong, 2017; Choi et al., 2018).

Addressing the persistent ethical and regulatory deficiencies, the subsequent chapter will systematically examine the methodology used to empirically assess their manifestation in real-world contexts. This will be accomplished through the implementation of a quantitative analysis of personality profiling techniques, complemented by a critical review of the purported methods Cambridge Analytica utilised in its psychographic profiling practices. Given that CA's specific methodologies are not publicly accessible due to the ICO's seizure of their servers and data (BBC, 2018b), the methodological framework will be developed through hypothesising. This approach aims to analyse the potential effects of psychographic profiling and analysis, to identify demographic groups that may be particularly susceptible to targeted advertising campaigns based on psychographic data.

Chapter 3:

25. Methodology Introduction

Understanding how CA conducts psychographic profiling and its methodologies remains undisclosed to the public. Notably, the ICO seized CA's servers and databases (BBC, 2018b). To critically examine how Facebook data and the OCEAN (Big Five) personality model are used, this paper formulates a hypothesis and employs a quantitative analysis framework. This involves creating a theoretical psychographic profile using publicly available datasets, such as the “Big Five Personality Test” from Kaggle (Tunguz, 2019), based on the OCEAN approach, and analysing various scoring metrics to understand the scoring process and associated personality traits. The approach supports the initial research objectives of investigating the influence and strategic effects of data-driven campaigns and establishing a comprehensive theoretical framework to assess potential adverse effects on targeted populations or issues. Quantitative analysis was chosen for its scalability, enabling the extraction of robust psychometric patterns from large datasets, which replicate real-world microtargeting strategies. Each methodology supports the paper's aim of ethically simulating data-driven microtargeting for risk and limitation evaluation (Bakir, 2020; McCrae and John, 1992).

26. Research Design

A quantitative approach is employed because the data are numerical, making this method appropriate. The dataset, from Kaggle's “Big Five Personality Test” (Tunguz, 2019), contains anonymised data from a large, diverse group of volunteers. It includes Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, which are standard in psychometric research (McCrae and John, 1992; Soto and John, 2017).

Since the research uses a quantitative approach, it aims to analyse patterns, traits, and distributions, including potential segmentations. It mimics psychometric methods used by CA. The dataset contains no personal information, and ethical standards are maintained.

The methodology uses statistical techniques, such as mean, standard deviation, and range, along with the k-means clustering algorithm, to identify trait patterns. This helps develop a psychographic segmentation to support potential microtargeting campaigns aligned with research goals.

Using secondary data sources for quantitative analysis eliminates the need to directly involve research participants, thus improving ethical compliance concerning informed consent and data confidentiality. This method also maintains the study's methodological validity within the ongoing dialogue on ethical standards in data science practices.

27. Data Source

The primary data utilised in this study was obtained from Kaggle, specifically the “Big Five Personality Test” dataset (Tunguz, 2019). This anonymised, secondary dataset comprises responses from an online personality assessment designed to evaluate the “Big Five” personality traits through the OCEAN framework. These dimensions are widely utilised in various research models (McCrae and John, 1992).

The sample was acquired from a broad, heterogeneous cohort of voluntary participants who completed the assessment online. This method allowed for a wide range of demographic and personality diversity within the sample. Participant responses were measured across five psychometric dimensions, with all personally identifiable information removed to ensure complete anonymity. This approach aligns with recognised ethical standards and research privacy regulations.

Since the sample was acquired from secondary sources, all data collection procedures for this study were performed independently, with no direct contact with participants. The dataset contains no identifiable information that could link the data to any individual participant. The dataset, sourced from Kaggle, is publicly accessible and has been anonymised to ensure compliance with ethical standards and data protection regulations. Consequently, this dataset is deemed suitable and reliable for the quantitative psychographic analysis presented in this study.

28. Data Preparation

The dataset was obtained from Kaggle as a .csv file and imported into Excel for cleaning and preprocessing. It was parsed to ensure each value was correctly placed in its cell. Missing values, duplicates, or invalid entries were identified and eliminated. The dataset contains 50,767,050 points and is 579.7 MB. Analysis will focus on UK participants' data, reducing volume and enabling targeted analysis. Data filtering used geographic metadata, which introduces sample bias and limits generalisability (Kekäläinen et al., 2020). It is also essential to verify that appropriate anonymisation is used to maintain ethical standards.

OCEAN dimension scores were verified for accuracy, recalculated if needed according to test guidelines. For large datasets, a randomised subset improves efficiency. Standardised variable naming ensures data integrity. This process upholds ethics and verifies data validity for analysis.

29. Analytical Methods

Data analysis will use Microsoft Excel for a quantitative study, with descriptive statistics (mean, max, min, range, SD) for the "Big Five" OCEAN traits to show personality distribution and variability.

Data visualisation employs histograms and boxplots to display trait score distributions and outliers. K-means offers interpretability and efficiency, identifying participant groups based on OCEAN traits. Hierarchical clustering was also considered (von Bergen and Diestel, 2025). The results should provide insights into trait patterns used for targeting, a method CA was accused of employing.

Since the dataset lacks gender stratification, formal hypothesis testing of variances won't be conducted. The analysis is limited to the UK, without city-level detail. Initial data analysis will use Microsoft Excel's statistical features for fundamental analysis and visualisation. Python, with pandas, matplotlib, and sklearn.cluster, will generate k-means clusters, simplifying the process (Arvai, 2020; Matplotlib, 2025; Pandas, 2025; W3Schools, n.d.).

Integrating descriptive, clustering, and relational methods enables a comprehensive, ethical investigation into psychometric assessment using the “Big Five” model in a UK sample.

30. Ethical Considerations

The methodology follows ethical standards for data science and human research. The dataset is anonymised, removing identifiable information such as names, locations, ages, and genders to protect confidentiality. Data were collected via voluntary online assessments without direct contact, so no additional ethical approval or active consent was required.

It is important to note that the dataset reports users' geolocation using latitude and longitude coordinates. According to Tunguz (2019), the dataset's author, these coordinates are approximate and not exact. Therefore, these coordinates will be excluded from the analysis during data sterilisation, as they do not significantly contribute to the study's accuracy.

GDPR compliance involves anonymising all data activities, with no personally identifiable information from the Kaggle dataset. This aligns with GDPR principles like data minimisation, purpose limitation, and security (ICO, 2023). The dataset is secondary, publicly accessible data, reducing privacy risks, making it ethically and legally suitable for research. However, using secondary anonymised data raises ethical concerns about participant consent, particularly when used to simulate industry practices (Markham and Buchanan, 2012).

Data management and analysis follow university protocols and the British Psychological Society's Code of Human Research Ethics, based on respect, competence, responsibility, and

integrity (British Psychological Society, 2021). These procedures protect privacy, foster transparency, and uphold research integrity responsibilities.

31. Limitations

This study has several limitations that warrant consideration. Firstly, the data used are from secondary, publicly accessible sources that are not curated by academic or research institutions, thus limiting control over data sample selection and potentially affecting data accuracy. Although the dataset is anonymised, ensuring that no personally identifiable information is included, certain variables, such as age, gender, socioeconomic status, and geographic location, restrict the ability to conduct detailed subgroup analyses.

The data was obtained from a large group of volunteers who completed an online assessment. This raises potential bias concerns, especially since responses were not moderated and relied on self-reporting, which may cause bias and misinterpretation. Psychographic segmentation categorises individuals by behaviour, but this dataset lacks that dimension, limiting understanding of how psychographic factors relate to targeting. As a result, microtargeting models may not fully reflect real-world ecological or behavioural effects (Prichard, 2021).

The dataset is analysed primarily in Microsoft Excel, though other software supporting advanced analysis is available. However, proficiency in these tools is often limited beyond Excel. These factors should be considered when interpreting results.

32. Summary

This chapter offers a comprehensive overview of a quantitative approach to analysing psychographic segmentation using secondary data from Kaggle's "Big Five Personality Test" dataset. The methodology includes data preprocessing, filtering for UK participants, and ensuring ethical compliance. It combines statistical analysis with k-means clustering aligned with the study's main aims. Analysing UK psychometric data provides rigorous, reproducible, and ethical insights. Established statistical and segmentation methods support exploring strategic and ethical issues in data-driven campaigns. The next chapter will present results, discussion, and strategies to combat misinformation on social media. References are based on key works in psychometrics, data ethics, and clustering to ensure a robust framework.

Chapter 4:

33. Results and Discussion: Introduction

This chapter provides a comprehensive critical analysis of the psychometric evaluation of data from UK-based participants, sourced from Kaggle’s “Big Five Personality Test” dataset, which uses the OCEAN framework. The quantitative methodology outlined in Chapter 3 aims to identify patterns in personality-driven targeting campaigns by applying the OCEAN psychographic profile, as used by the data science company Cambridge Analytica.

The chapter begins with a comprehensive statistical analysis of the dataset's variability, which will be systematically summarised and presented. Next, a psychographic segmentation will be conducted using k-means clustering. The discussion will then interpret these findings within the context of existing academic literature, emphasising their strategic and ethical implications while acknowledging potential limitations. Considering concerns about psychometric targeting and the spread of misinformation, a mitigation strategy will be proposed in the form of a Python script outlining a theoretical and practical intervention to reduce the dissemination of misinformation on social media platforms.

This chapter aims to provide a thorough exploration of the practical applications, strategic issues, and ethical considerations related to psychometric profiling in digital data science.

34. Descriptive Statistics

This section presents statistical analyses of each OCEAN personality trait for participants based in the UK, utilising data obtained from Kaggle’s (Tunguz, 2019). The dataset, comprising 50,767,050 data values from a global participant pool, has been filtered to include only individuals based in the UK. This refinement not only improves the manageability of the dataset but also aligns with the intention to simulate targeted advertising, considering the Cambridge Analytica scandal.

For the initial analysis, a comprehensive overview of the statistical data extracted from the “Big Five” dataset was conducted. Although the dataset is confined to participants residing in the UK, it includes a substantial number of entries, 66,487 per trait, totalling 3,324,350 data values.

Please see Appendix No. 1: Kaggle Dataset for OCEAN traits for UK participants.

| Traits | Mean | SD | Min | Max | Range |
|-------------------|-----------|-----------|-----|-----|-------|
| Openness | 3.0015326 | 1.3168078 | 0 | 5 | 5 |
| Conscientiousness | 3.0947012 | 1.324662 | 0 | 5 | 5 |
| Extraversion | 3.122275 | 1.3765079 | 0 | 5 | 5 |
| Agreeableness | 3.074956 | 1.3029307 | 0 | 5 | 5 |
| Neuroticism | 3.2867282 | 1.3908787 | 0 | 5 | 5 |

Table 1: OCEAN traits illustrating the data distribution (Microsoft Excel, 2024).

The analysis uses statistical measures such as mean, SD, Min, Max, and range to summarise the data distribution. These descriptors are essential for exploratory data analysis (EDA), a common practice in fields such as psychological research. (Data Science Discovery, n.d.; Hayes, 2024; Kekäläinen et al., 2020).

Table 1 presents the five traits of the OCEAN psychographic profile, along with their statistical parameters: Mean, SD, Min, Max, and Range. Min and Max are scaled 0-5, with the Range always 5. Variations mainly appear in the Mean, which centres around 3, and the SDs of about 1.3 and 1.4, indicating moderate trait variability.

The Neuroticism trait exhibits a mean score of $m = 3.28$ with an $SD = 1.39$, indicating that participants tend to score relatively higher on this trait in comparison to other traits.

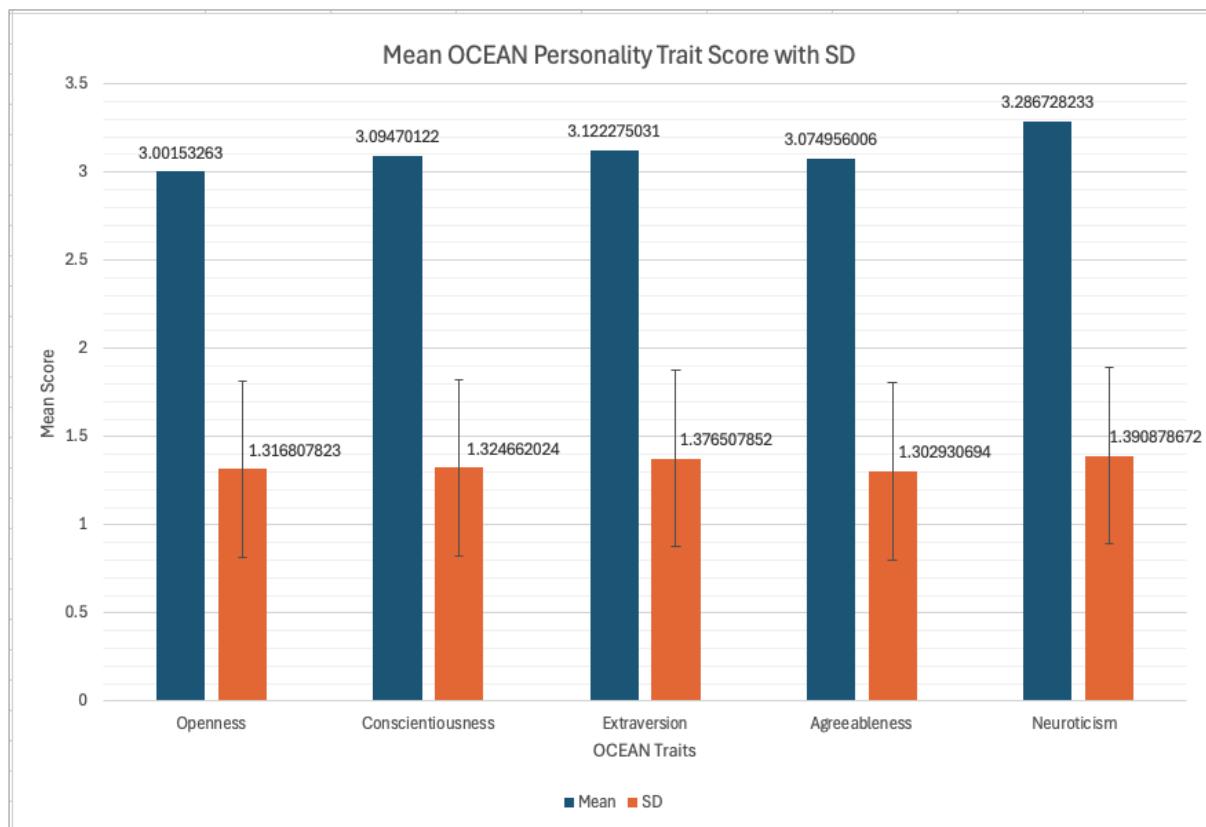


Figure 2: A bar chart with error bars displays the Mean and SD for each trait, enabling easy comparison of average levels and variability across all five traits. It visually compares each trait using error bars to highlight trait variability (Microsoft Excel, 2024).

Figure 2 presents data from Table 1, with blue bars indicating means and orange bars representing standard deviations. Vertical error bars show trait variability, facilitating visual interpretation. The x-axis displays the OCEAN model, and the y-axis depicts mean scores, illustrating both central tendency and trait variability.

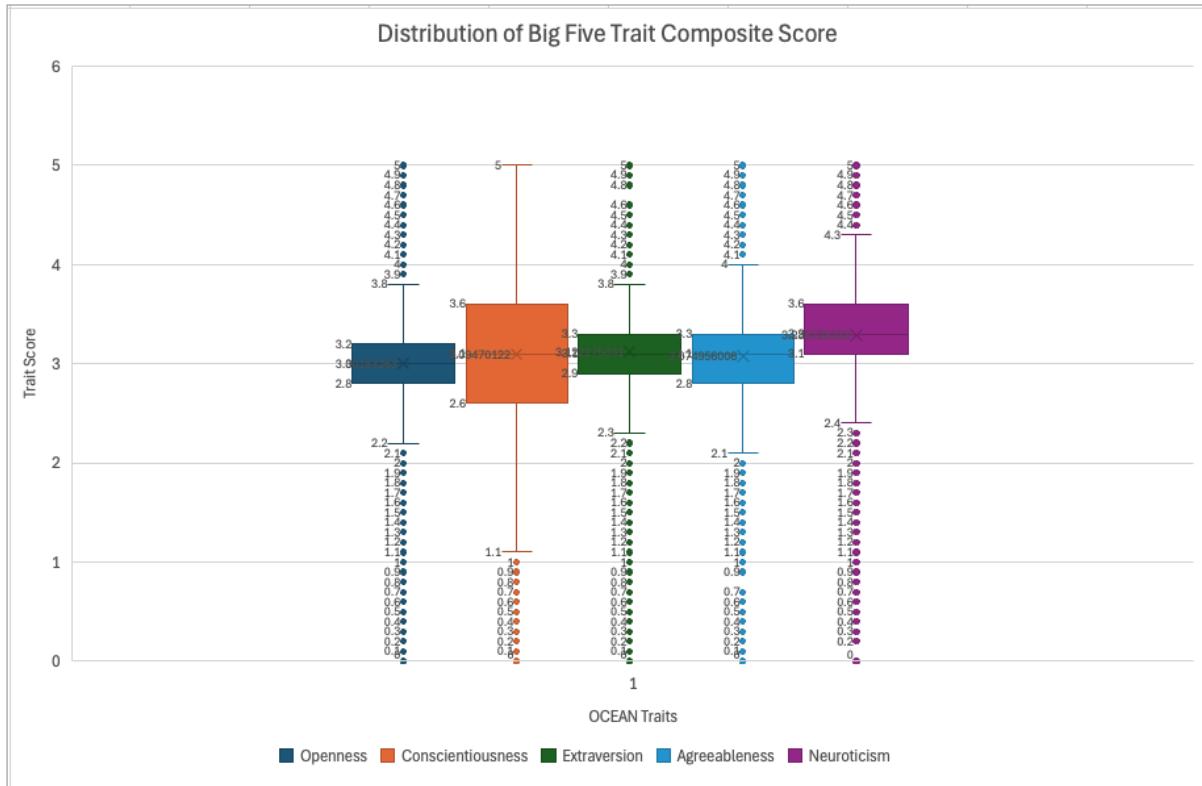


Figure 3: A Boxplot graph that displays all five OCEAN traits, showing the interquartile range, median, and outliers for each trait, confirms that the population's central tendency and spread are consistent across all OCEAN dimensions (Microsoft Excel, 2024).

The Boxplot in Figure 3 visually summarises the distribution, central tendency, variability, and outliers for each OCEAN trait. The median, shown by the central line, indicates the central score. The interquartile range, marked by the box edges, shows the middle 50%. Whiskers extend to the minimum and maximum non-outlier values, with outliers represented as dots beyond these ranges. The mean score, representing the average, offers additional insight into the distribution.

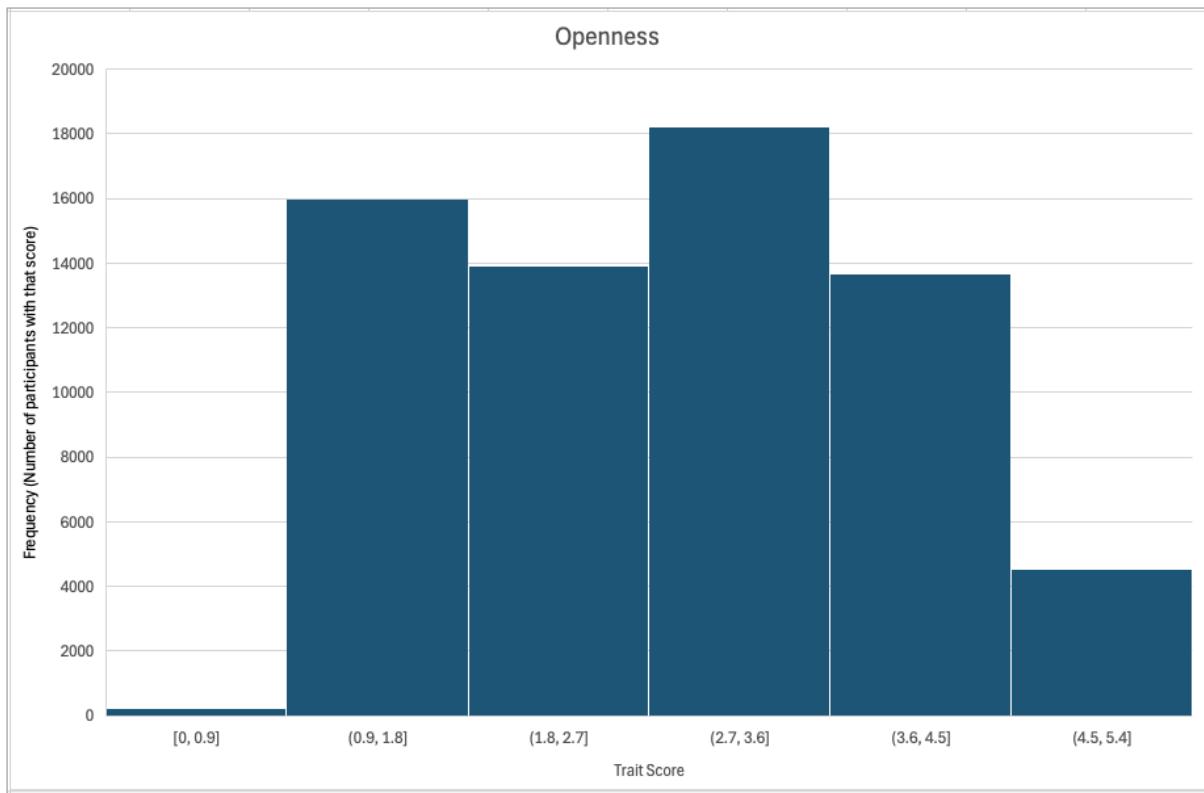


Figure 4: Histogram for Openness trait (Microsoft Excel, 2024). The distribution demonstrates a balanced and uniform spread, characterised by a subtle concentration within the central range, specifically between 2.7 and 3.6. Based on the histogram, participants exhibit moderate levels of openness (Hill and Edmonds, 2017; Soto and John, 2017; von Bergen and Diestel, 2025).

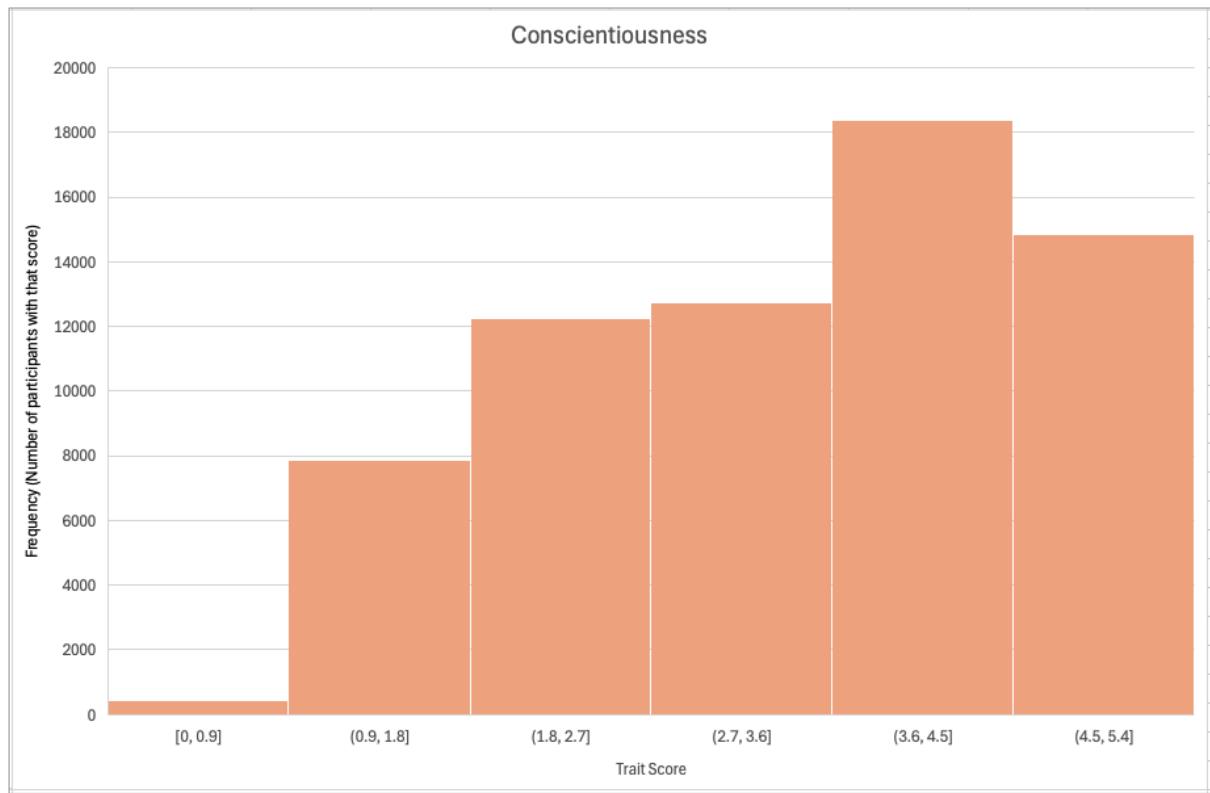


Figure 5: Histogram for Conscientiousness trait (Microsoft Excel, 2024). Most participants' scores clustered within the upper-middle range, specifically between 2.7 and 4.5. The distribution skewed towards the higher end of the spectrum, suggesting that participants tend to be responsible (Hill and Edmonds, 2017; Soto and John, 2017; von Bergen and Diestel, 2025).

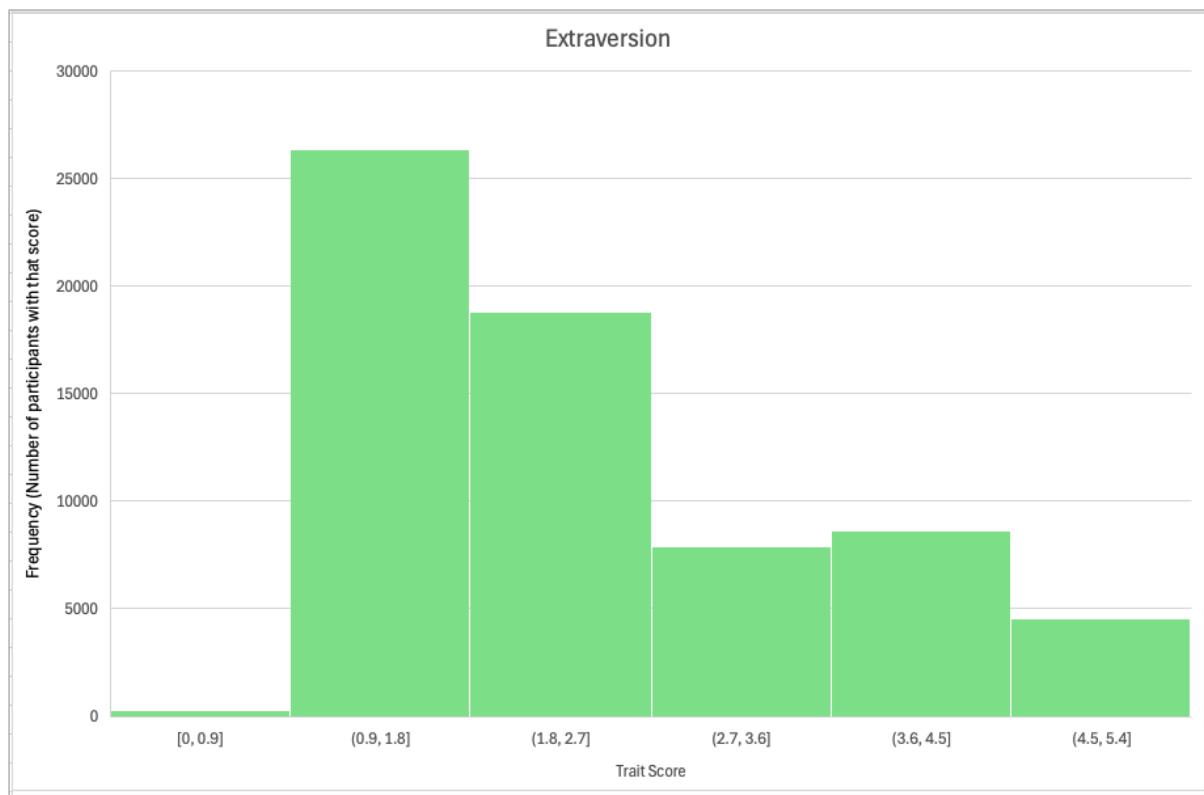


Figure 6: Histogram for the Extraversion trait (Microsoft Excel, 2024). The distribution exhibits pronounced left skewness, with the data points concentrated around 0.9 and 3.6. This indicates that the participant may exhibit a tendency toward social withdrawal and a reduced propensity for social engagement (Hill and Edmonds, 2017; Soto and John, 2017; von Bergen and Diestel, 2025).

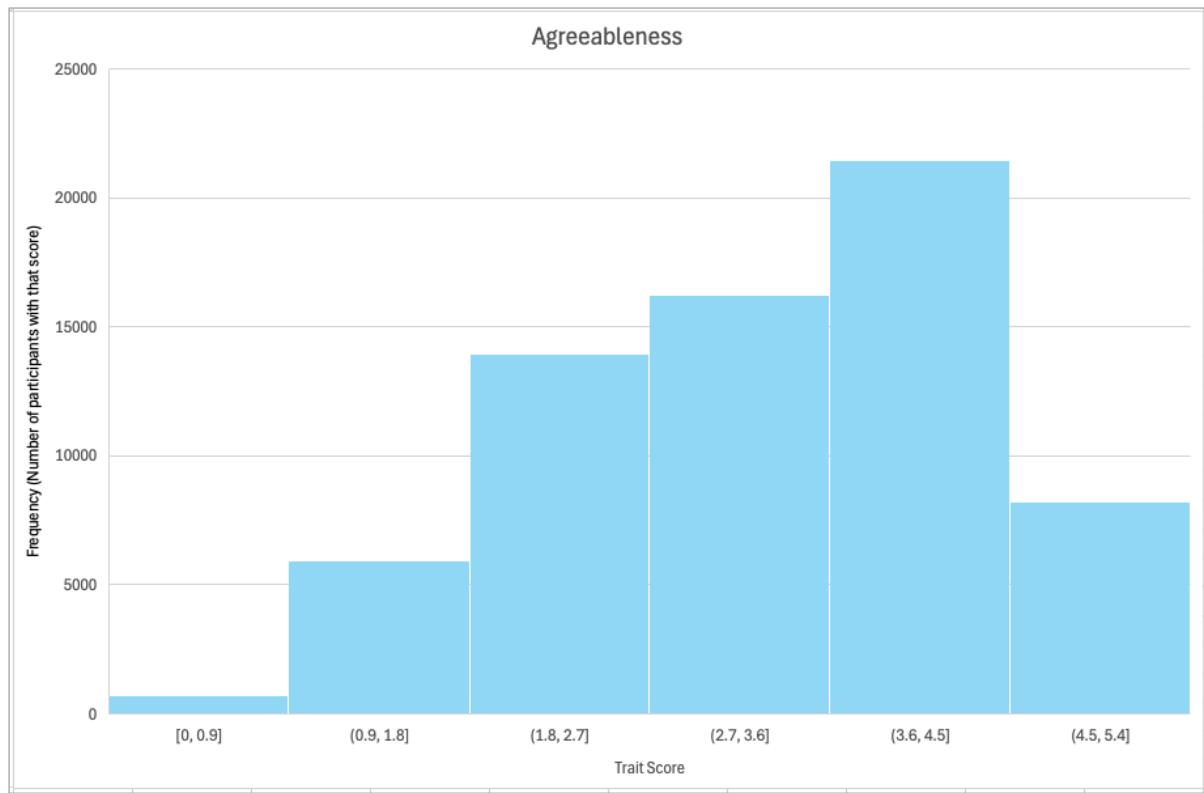


Figure 7: Histogram for the Agreeableness trait (Microsoft Excel, 2024). The distribution exhibits a right-skewed pattern, characterised by a higher concentration of values within the mid to upper ranges, specifically between 2.7 and 4.5. This suggests that most individuals demonstrate moderate to high levels of agreeableness (Hill and Edmonds, 2017; Soto and John, 2017; von Bergen and Diestel, 2025).

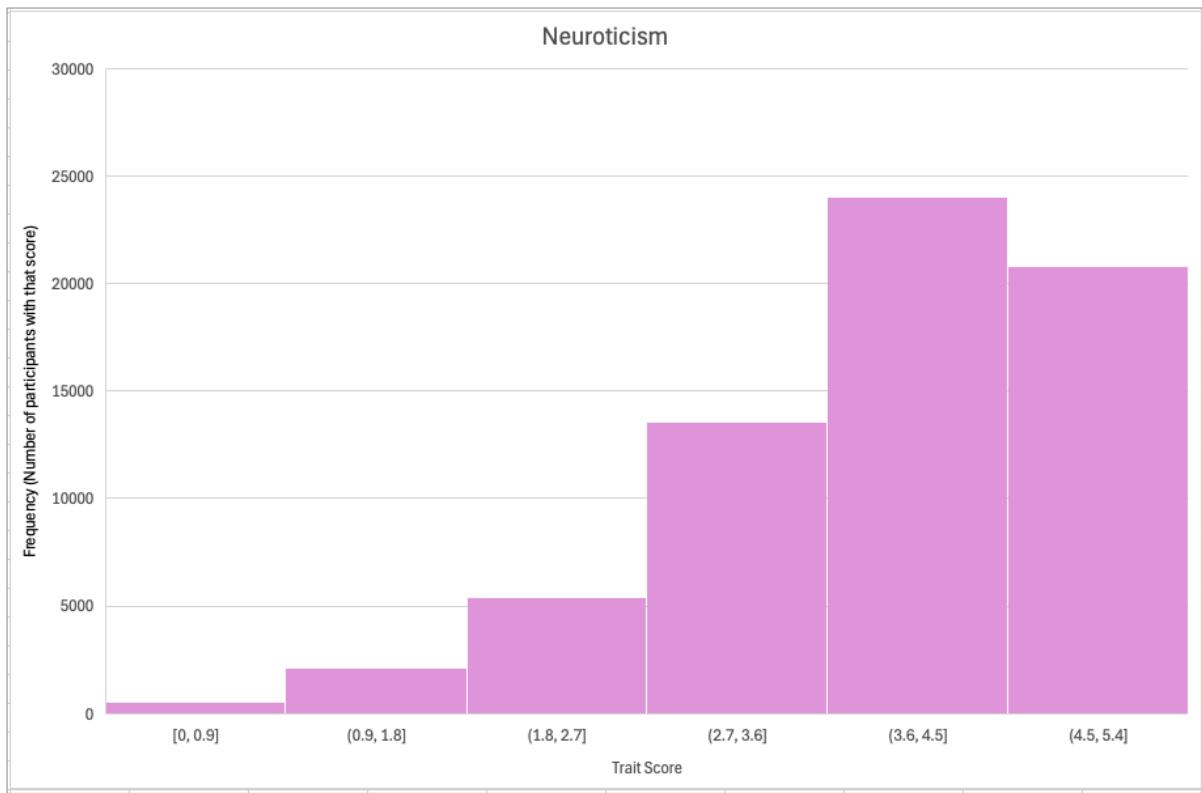


Figure 8: Histogram for Neuroticism trait (Microsoft Excel, 2024). The histogram exhibits right skewness, with a pronounced concentration of observations in the upper bin, ranging from 3.6 to 5.4. Elevated scores within this interval may indicate a higher level of neuroticism among participants (Hill and Edmonds, 2017; Soto and John, 2017; von Bergen and Diestel, 2025).

Figures 4 through 8 present histograms of the various distributions associated with the five individual OCEAN traits. Here are some of the observations

- Openness is centrally distributed.
- Conscientiousness is skewed higher, which may suggest responsible behaviour.
- Extraversion is more left-skewed, indicating a preference for introversion.
- Agreeableness is skewed to the right, suggesting they are quite social.
- Neuroticism is skewed to the right, suggesting that participants have high levels of neuroticism.

Table 2 provides a comprehensive overview of these histograms, synthesised from the observations in Figures 4 to 8.

| Histogram Summary | | | | |
|-------------------|------------------|-------------|-----------------|---------------------------|
| Trait | Distribution | Trait Range | Spread | Interpretation |
| Openness | Moderate - hight | 2.7 - 3.6 | Fairly Balanced | Open to new ideas |
| Conscientiousness | High | 3.6 - 4.5 | Right | Responsible and organised |
| Extraversion | Low - moderate | 0.9 - 2.7 | Left | Reserved and introverted |
| Agreeableness | Moderate - hight | 2.7 - 4.5 | Right | Friendly and cooperative |
| Neuroticism | High | 3.6 - 5.4 | Right | Emotional and sensitive |

Table 2: A summary of the histogram graph based on the five individual OCEAN traits, illustrating the psychometric diversity for clustering and potentially targeting (Lim, 2025; Microsoft Excel, 2024; Sutton, 2025).

35. Cluster Analysis

A k-means clustering process was implemented to delineate underlying psychographic segments among participants. This involved grouping individuals based on their personality profiles derived from aggregating OCEAN traits, thereby facilitating targeted outreach strategies analogous to those used by organisations such as Cambridge Analytica.

For the cluster analysis, a cluster number of seven ($k = 7$) was selected, as this configuration yielded sufficient data to facilitate an initial examination and identify potential patterns in the trait that could inform micro-targeted advertising strategies.

The cluster analysis utilised Python alongside pandas, matplotlib, and sklearn.cluster libraries to develop a k-means model. Google Colab served as the IDE with autofill features (Codecademy, 2019). The dataset, limited to UK data and five OCEAN traits, excluded other variables. The filtered data was stored as CSV files. Data for each trait was averaged with (=AVERAGE(xx:xx)) for the K-means clustering (Arvai, 2020; GeeksforGeeks, 2021; Google Colab, 2019; Matplotlib, 2025; Pandas, 2025; Scikit-learn, 2019; W3Schools, n.d.; W3Schools, 2024; W3Schools, 2025).

Please see Appendix No. 2: K-Means Python Code, which includes both the .csv file and Colab Python code.

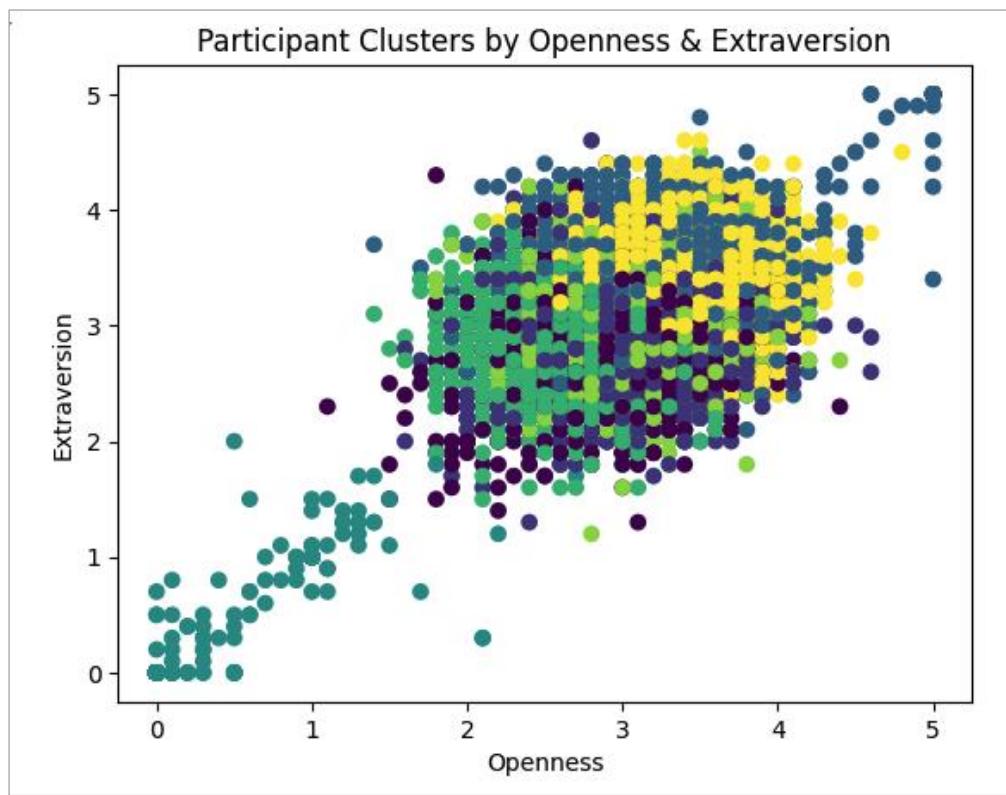


Figure 9: Cluster from UK participant based on Openness and Extraversion (Google Colab, 2019).

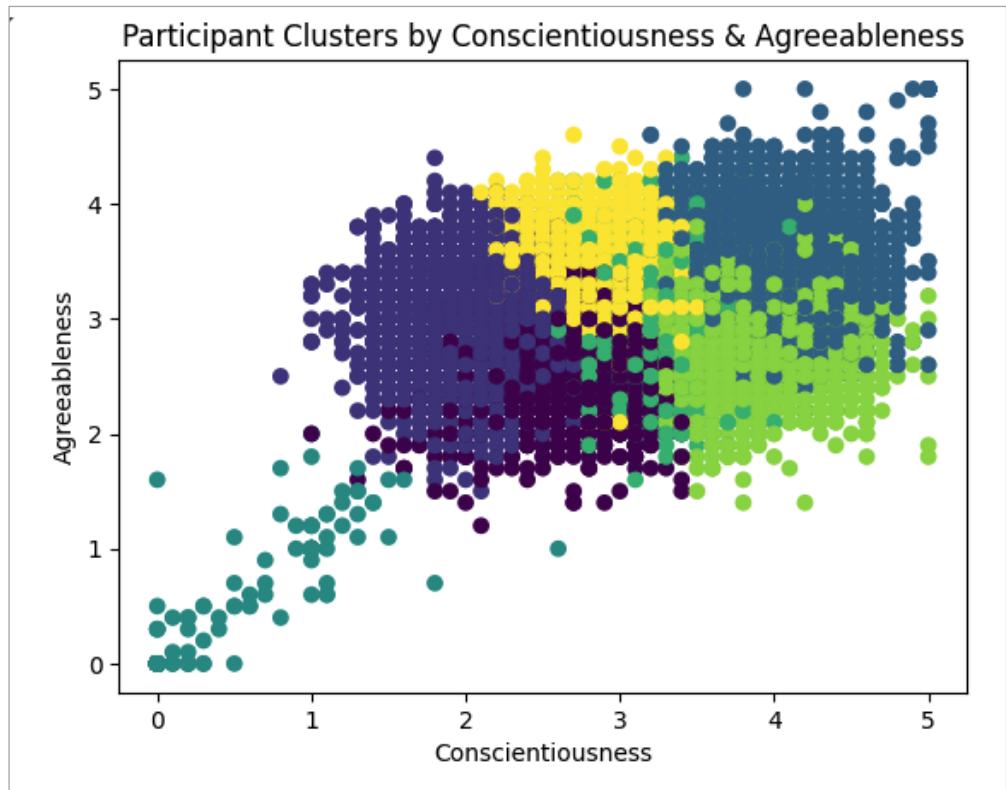


Figure 10: Cluster from UK participant based on Conscientiousness and Agreeableness (Google Colab, 2019).

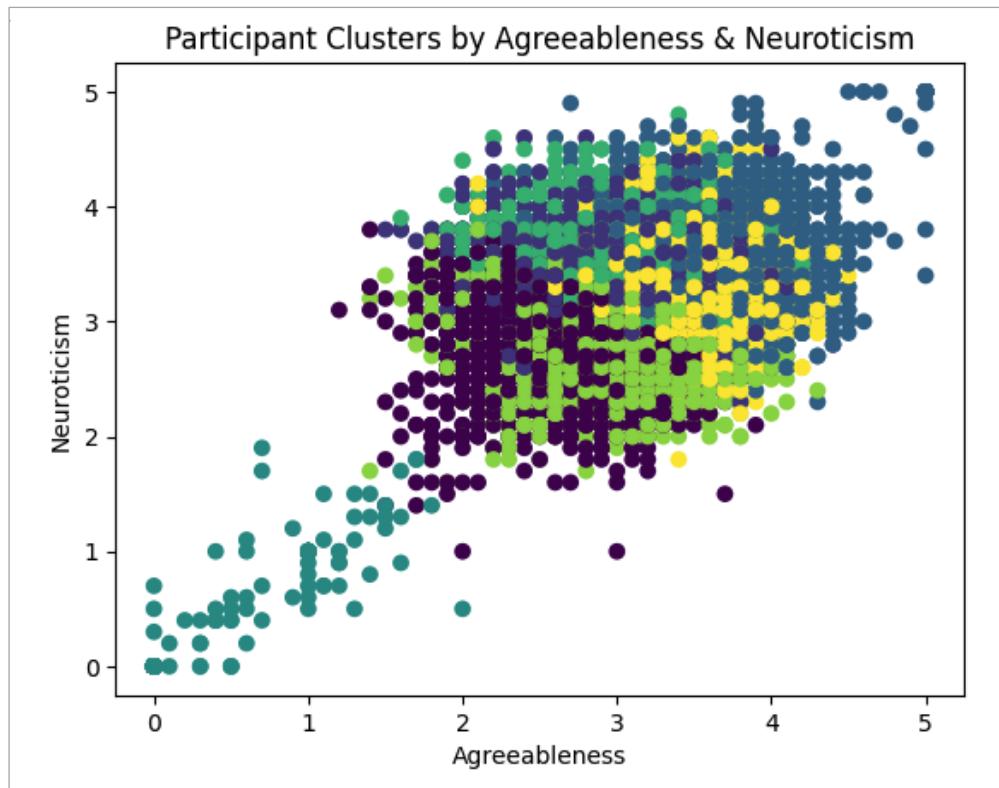


Figure 11: Cluster from UK participant based on Agreeableness and Neuroticism (Google Colab, 2019).

Each cluster example indicates a central pattern, suggesting that the participant exhibits a similar configuration across these instances.

| Cluster centres (average trait scores per group): | | | | | |
|---|----------|-------------------|--------------|---------------|-------------|
| No: | Openness | Conscientiousness | Extraversion | Agreeableness | Neuroticism |
| 0 | 2.994491 | 2.742674 | 2.982321 | 2.772797 | 2.92775 |
| 1 | 3.008246 | 2.210734 | 3.050879 | 3.014466 | 3.413152 |
| 2 | 3.063043 | 3.887495 | 3.336722 | 3.418302 | 3.531913 |
| 3 | 0.361607 | 0.322768 | 0.345982 | 0.336161 | 0.311161 |
| 4 | 2.781251 | 3.22075 | 2.967832 | 3.062127 | 3.517167 |
| 5 | 2.997422 | 3.689652 | 3.117689 | 2.956861 | 2.989889 |
| 6 | 3.215027 | 2.954407 | 3.333698 | 3.271563 | 3.369383 |

Table 3: Cluster centres table based on average trait from each of the OCEAN groups (Google Colab, 2019; Microsoft Excel, 2024).

Based on the mean value from Table 1, which falls within the range $m = 3.0$ to $m = 3.2$, it can be inferred that all traits are within the designated target parameters, indicating a state of equilibrium or balance. Consequently, these traits demonstrate the potential for micro-targeting.

Observe the cluster comprising rows 1, 4, and 6, where the Neuroticism scores are comparatively elevated, ranging from 3.3 to 3.5. These scores are significantly higher than those observed in other clusters and across different traits.

To comprehend the traits, consider their foundational model, though a detailed review is beyond this paper's scope. Many scholars have examined this area, positioning traits along various continuums. This approach would help interpret Table 3 and the participant cluster's placement within that spectrum.

Please see Appendix No. 3: Table 4 of OCEAN Traits, including their score values and psychological profile descriptions.

Table 4: Provides a succinct overview of personality traits based on their high and low score values, which collectively outline an individual's psychological profile (Lim, 2025; Nasello et al., 2023; Sutton, 2025).

36. Discussion

To achieve a comprehensive understanding of which traits are susceptible to micro-targeting, Table 5 displays the mean and SD for each OCEAN personality dimension, excluding minimum, maximum, and range, since these are constant. It also includes a histogram of trait distribution, matched with psychological data from Table 4 to classify participants as high or low on specific trait characteristics.

| Trait | Mean | SD | Distribution and trait range | Interpretation | Susceptibility to microtargeting ads |
|-------------------|------|------|-------------------------------------|---|--------------------------------------|
| Openness | 3.00 | 1.32 | Fairly Balanced (2.7 – 3.6) | Tolerant, imaginative, emotionally sensitive, independent | Moderate |
| Conscientiousness | 3.09 | 1.32 | Right - High (3.6 – 4.5) | Self-disciplined, motivated, organised, and considerate | Moderate – High |
| Extraversion | 3.12 | 1.38 | Left – Low / Moderate (0.9 – 2.7) | Sociable, enthusiastic, and assertive | Moderate |
| Agreeableness | 3.07 | 1.30 | Right – Moderate / High (2.7 – 4.4) | Modest, trustful, cooperative, | Moderate |

| | | | | | |
|-------------|------|------|-----------------------------|--|------|
| | | | | and straightforward | |
| Neuroticism | 3.29 | 1.39 | Right - High (3.6 – 5.4) | Hostile, angry, impulsive, anxious, and self-conscious | High |

Table 5: An overview based on the data summary collected from the data distribution in Table 1, a summary from the histogram in Table 2, and the individual trait psychological profiling information in Table 4 (Lim, 2025; Sutton, 2025).

The distribution presented in Table 5 pertains to UK participants. It suggests that the sample shows moderate to high susceptibility to traits relevant to microtargeted advertising, with Neuroticism demonstrating the highest susceptibility.

| No: | Cluster centres (average trait scores per group): | | | | |
|-----|---|-------------------|--------------|---------------|-------------|
| | Openness | Conscientiousness | Extraversion | Agreeableness | Neuroticism |
| 0 | 2.994491 | 2.742674 | 2.982321 | 2.772797 | 2.92775 |
| 1 | 3.008246 | 2.210734 | 3.050879 | 3.014466 | 3.413152 |
| 2 | 3.063043 | 3.887495 | 3.336722 | 3.418302 | 3.531913 |
| 3 | 0.361607 | 0.322768 | 0.345982 | 0.336161 | 0.311161 |
| 4 | 2.781251 | 3.22075 | 2.967832 | 3.062127 | 3.517167 |
| 5 | 2.997422 | 3.689652 | 3.117689 | 2.956861 | 2.989889 |
| 6 | 3.215027 | 2.954407 | 3.333698 | 3.271563 | 3.369383 |

Table 3: Cluster centres table based on average trait from each of the OCEAN groups, with highlighted rows 1, 4, and 6, indicating the high levels of Neuroticism (Google Colab, 2019; Microsoft Excel, 2024).

Neuroticism is the most susceptible OCEAN trait to microtargeting, with a mean of $m = 3.29$ and $SD = 1.39$, the highest among all traits in Table 1. People high in Neuroticism tend to be hostile, anxious, impulsive, and self-conscious. Extraversion ranks second, with $m = 3.12$ and $SD = 1.38$, and is characterised by sociability, enthusiasm, assertiveness, and warmth, indicating moderate susceptibility. These results highlight differences in vulnerability across personality traits in targeted marketing strategies (Lim, 2025; Sutton, 2025).

In Table 3, rows 1, 4, and 6, participants show high Neuroticism with a right-skewed distribution, suggesting susceptibility to microtargeting. Data science companies like Cambridge Analytica could exploit this by customising ads to influence fears and anxieties. Such a hypothesis is corroborated by Bakir (2020), who states that those high in neuroticism are more vulnerable to fear messages. When combined with geographic data, Cambridge Analytica might have used microtargeting based on neuroticism profiles and location to evoke fear responses. This assertion is further supported by Prichard (2021), who confirms that Cambridge Analytica employed unethically obtained personal data for microtargeting, thereby increasing click rates (Resnick, 2018).

Chapter 2 highlights key ethical issues, notably Cambridge Analytica's unconsented data collection and concealment via SCL Group. It targeted individuals using psychological profiles to spread misinformation, risking physical conflict and biased judgments on race, religion, beliefs, and sexual orientation.

The Southport Stabbing on 29th July 2024 involved multiple assaults at a dance studio. Initially, reports falsely claimed a Muslim asylum seeker, "Ali-Al-Shakati," entered the UK illegally and was the attacker. This misinformation spread widely on platforms like X, Facebook, Telegram, and Instagram, with X's coverage reaching around 1.5 million views, amplified by far-right influencers, including Andrew Tate's post with over 9 million followers, believed to reach over 15 million. This false information incited violence, riots, attacks on religious sites, and vandalism. Later, investigations revealed the attacker was a British national, "Axel Rudakubana," with no links to Islam or illegal immigration. (BBC Bitesize, 2024; Fung, 2024; Kiderlin, 2024; Mohamed, 2024; Shah, 2024).

This excerpt emphasises misinformation's role, spread via social media, which heightened tensions and resulted in violence. Mohamed (2024) states that it received about 27 million impressions, demonstrating its broad influence. Alarmingly, social platforms initially failed to address it. Facebook's AI promotes popular content, prioritising profits over safety, indicating a focus on financial gain rather than user protection (Benesch, 2021).

37. Mitigation Strategy

The proliferation of misinformation on social media remains a significant concern. The UK introduced the Online Safety Act 2023 to regulate harmful content (Legislation.Gov.UK, 2023). However, Woods (2024) notes that the legislation does not explicitly target misinformation or disinformation, as it lacks criteria for criminality or deliberate harm.

To address these concerns, a hypothetical software application could be integrated into social media platforms such as Facebook to identify misinformation, based on UK government initiatives aimed at combating misinformation by cross-referencing advertisements and viral content using a news API. While demonstrating potential, deploying it faces challenges due to opaque platform mechanisms and the need for platform and government cooperation intervention.

Despite implementation challenges, this approach can mitigate misinformation by informing users who engage with content about its verification status. If supported by the news API, it provides relevant sources. Alternatively, it assists users in classifying content as factual or potentially misleading.

Please see Appendix No. 4: Mitigation Hypothesis Code, which includes a README file and code snippets.

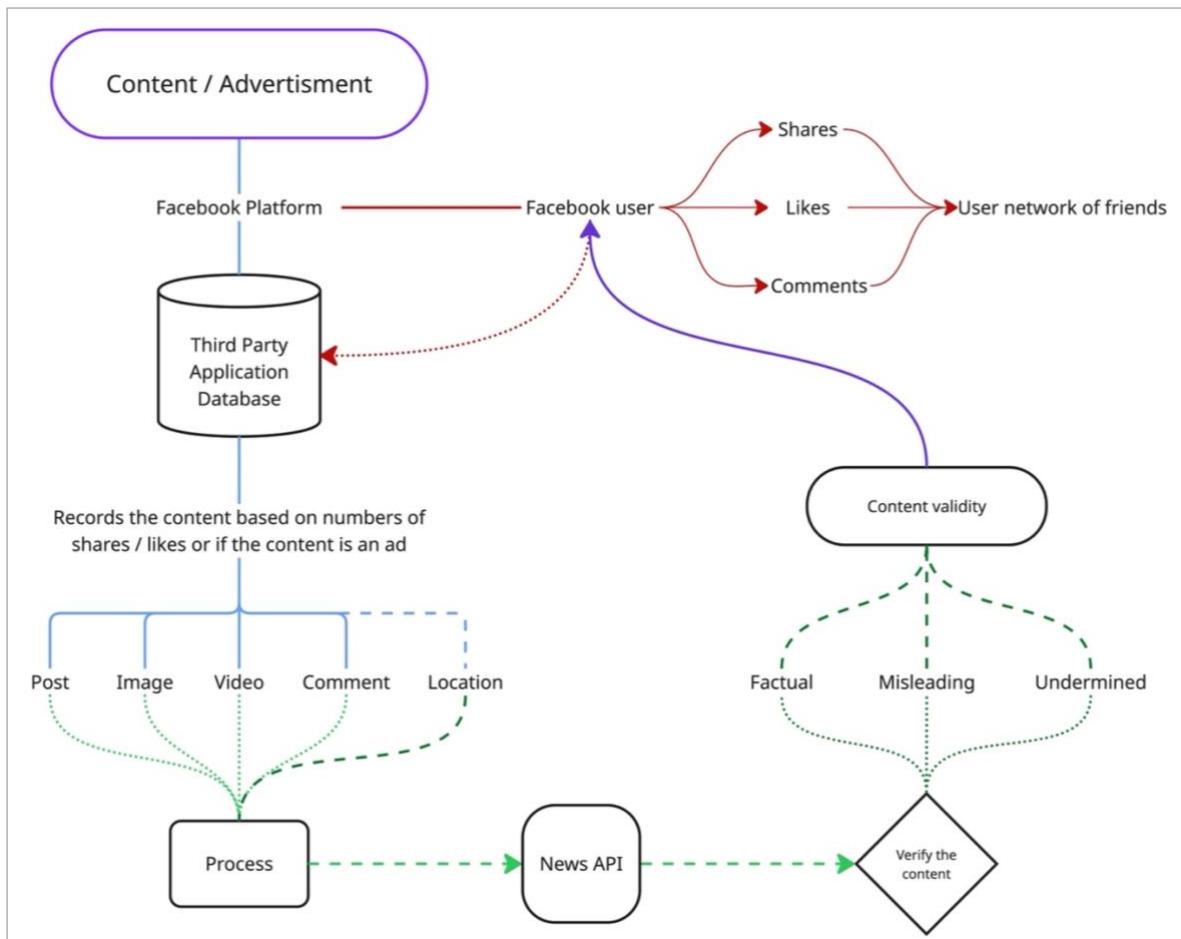


Figure 12: The system flowchart provides an overview of the application's functionality.

The system flow diagram, Figure 12, illustrates the process by which, upon content being presented on the Facebook platform, the associated application, with its own database for categorising relevant information, analyses user interactions such as shares, likes, and comments. Subsequently, the content is validated by cross-referencing it with a news API to confirm its authenticity. Once verified, the system offers users additional information to support the content's veracity and reliability.

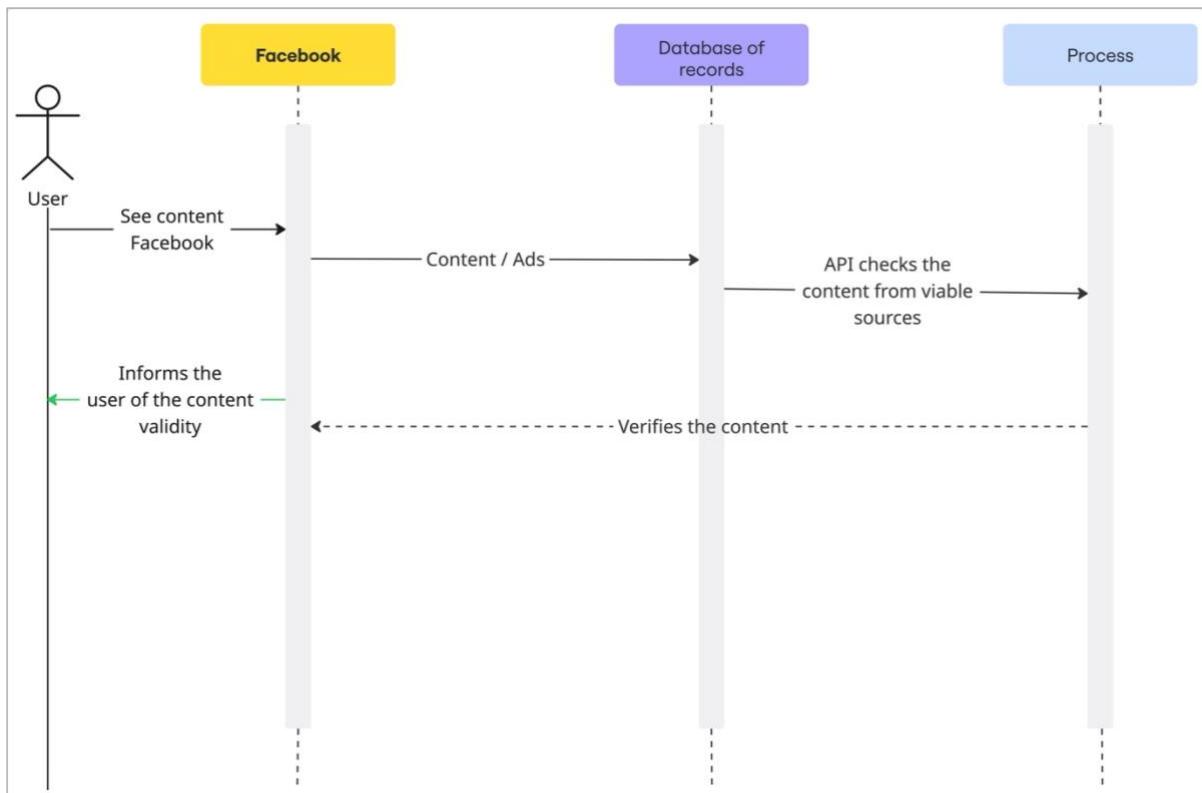


Figure 13: UML Sequence diagram, illustrating the user's journey through accessing content on Facebook.

The illustration in Figure 13 offers a comprehensive overview of user interactions with content within the Facebook platform. Upon engagement with a specific piece of content, the data is transmitted to an external database, which operates independently of Facebook's proprietary system. This external database performs a validation process to assess the authenticity of the content, then relays the results to the user.

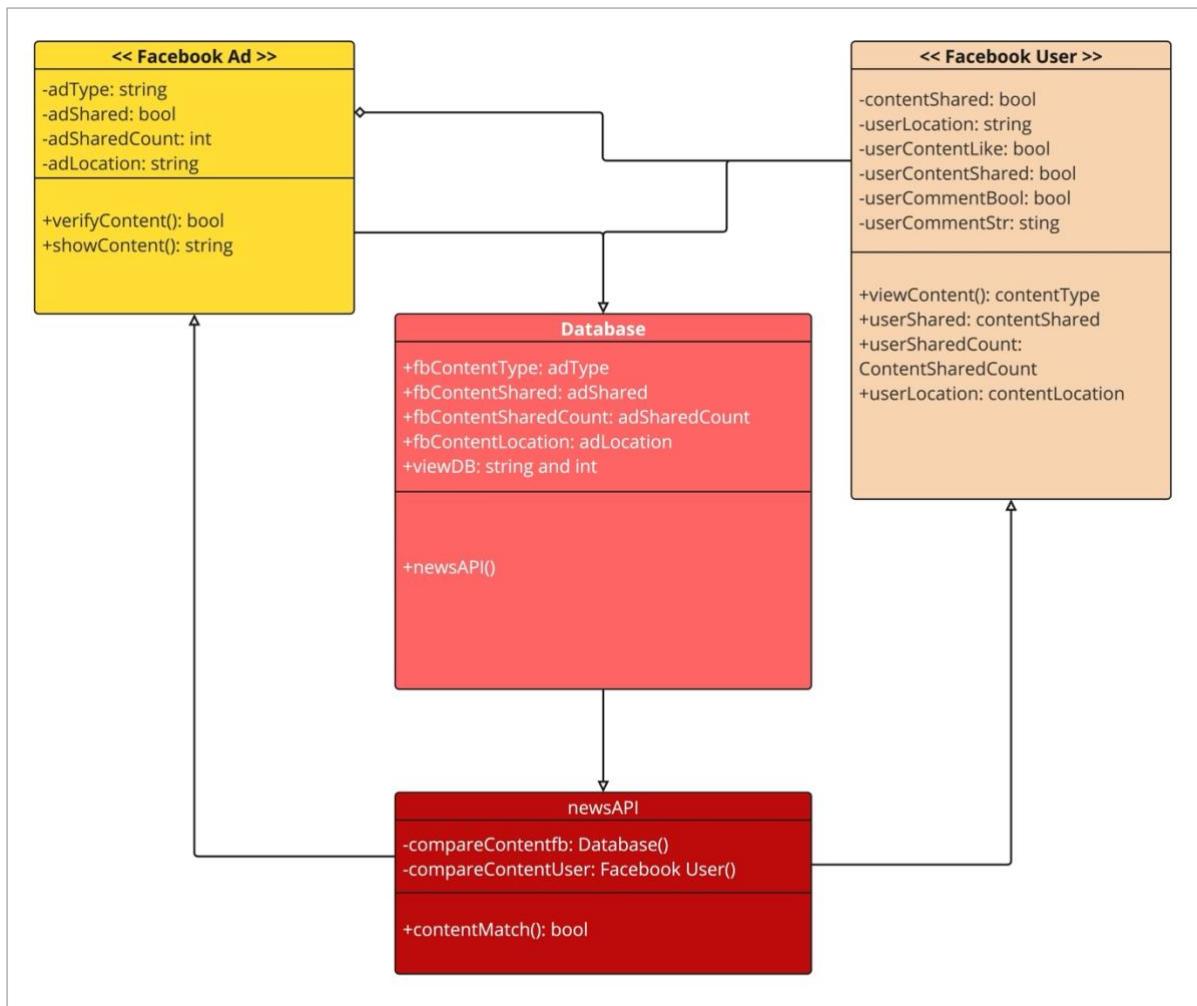


Figure 14: The UML Class diagram provides a high-level overview of the application system.

The UML class separates ads and users, as general users rarely create targeted ads, which corporations mainly use. Content data is stored externally, independent of Facebook, to match, store, and connect to a news API. The system will verify content authenticity to raise user awareness.

38. Summary

This chapter presents empirical studies on UK participants' OCEAN traits and their susceptibility to microtargeted advertising, mainly misinformation. Data from Kaggle supports ethical and valid research (Tunguz, 2019). Analysis included descriptive stats and visualisations such as bar charts, histograms, and boxplots. K-means clustering identified psychographic segments, with neuroticism showing the highest susceptibility, likely due to emotional traits and fear vulnerability. Visual data confirm that many participants are moderately to highly susceptible across personality traits.

Findings suggest that psychometric profiling is a strategic tool for targeted messaging, like Cambridge Analytica. The study discusses the benefits and ethical issues of using personal data for content targeting, including privacy and misinformation concerns.

A conceptual approach to enhance awareness and promote responsible data use is proposed to combat misinformation. Recommendations include improved oversight, greater transparency, and further research for ethical management of social platform data.

Chapter 5:

39. Research Aim and Methods

This study aims to examine the ethical obligations of data science companies critically and to scrutinise the methodologies employed in their unethical data collection practices.

Specifically, the research investigates how harvested data is utilised to create psychometric profiles for microtargeting Facebook users with tailored advertisements, which are inherently manipulative and designed to sway user behaviour in accordance with the strategic objectives of the data science entity. The analysis further explores the relationship between the “Big Five” personality traits (OCEAN) among a sample of UK participants and their susceptibility to microtargeting, leveraging an extensive Kaggle dataset. Methodologically, the study integrates statistical testing through hypothesis formation grounded in the data practices of the companies examined, with the results visualised to highlight psychometric patterns that may have contributed to user vulnerability (Lim, 2025; Sutton, 2025; Tunguz, 2019).

40. Main Findings

The analysis, derived from a cluster analysis of UK participants, indicates that individuals exhibiting the Neuroticism psychographic trait demonstrate increased susceptibility to microtargeted advertising campaigns, primarily due to elevated levels of emotional fear. Consequently, fear appeals are likely to provoke behavioural responses within this group. These findings align with existing research suggesting that individuals with high Neuroticism are more vulnerable to influence and manipulation through digital advertising, as supported by studies (Bakir, 2020; Prichard, 2021). Furthermore, visual data analysis reveals that a substantial subset of the population displays moderate to high levels across all measured traits, suggesting that psychometric targeting, as employed by Cambridge Analytica, may exert broad influence across various personality dimensions.

41. Implications

The use of psychometric profiling offers a powerful tool to tailor digital content to elicit a response and has demonstrated its effectiveness in politics and commerce (Bakir, 2020; Resnick, 2018). Despite the emergence of psychometric profiling for targeting campaigns and its technical advances, the methodology has raised several ethical concerns and challenges, as large-scale personality profiling may blur the line between persuasion and manipulation, alternatively changing the behaviour of the targeted individual by polluting their thoughts with misinformation (Cadwalladr and Graham-Harrison, 2018; BBC, 2018a; Hinds et al., 2020). The digital landscape, with its opaque algorithms and weak regulatory framework, has created the ideal environment for abuse and public harm; recent incidents involving

misinformation that led to civil unrest exemplify this (BBC Bitesize, 2024; Fung, 2024; Mohamed, 2024).

42. Ethics

The paper delineates several ethical concerns, including the unauthorised utilisation of personal data without explicit informed consent, the potential for algorithmic biases that may facilitate manipulation, and the dissemination of misinformation capable of eroding social trust and stability (British Psychological Society, 2021; ICO, 2023). Data science companies and digital institutions whose primary business model relies on user data, like social media platforms such as Facebook, hold considerable sway over user experience and have increasingly blurred the line between data-driven services and digital exploitation (Benesch, 2021; Cadwalladr, 2019). Moreover, regulatory frameworks and government agencies often struggle to keep up with technological advances, leaving the complexities and scale of psychometric targeting insufficiently addressed. The use of such techniques in political settings presents serious risks to democratic processes. Therefore, it is essential to strengthen ethical oversight, improve user protections, increase transparency of data-driven organisations, and advance the development of explainable artificial intelligence systems (Afsharian, 2025; Dowling, 2022; Markham and Buchanan, 2012).

43. Limitations

The research predominantly relies on secondary data sources from Kaggle, which are publicly accessible but not curated or verified by scholarly or research institutions. This reliance imposes limitations on controlling sample selection and may affect data accuracy, potentially introduce bias, and reduce the relevance of findings to real-world settings, thus restricting the scope for thorough testing. While the dataset has been anonymised to address ethical issues and ensure research validity, it excludes identifiable demographic variables, such as age, gender, socioeconomic status, and geographic location, limiting the capacity for detailed subgroup analyses. Psychographic segmentation categories are built on behavioural indicators; however, the dataset lacks this specific information, reducing insights into the relationship between psychographics and targeting strategies. Consequently, models for microtargeting derived from this data may not fully capture actual campaign behaviours or ecological validity (Barth and de Jong, 2017; Prichard, 2021; Tunguz, 2019). Furthermore, the data sampling was confined to the UK, raising concerns about whether the findings can be applied to regions outside major social media markets, thereby introducing potential regional bias.

44. Future Research

Future research in this domain should endeavour to replicate existing findings across diverse, multinational cohorts. It is essential to utilise regions comparable to those involved in the Cambridge Analytica case to elucidate the clustering traits derived from psychometric profiling. Data collection should be conducted independently of secondary sources to ensure academic integrity and authenticity. Additionally, such studies should assess the effectiveness of mitigation strategies implemented on live platforms and consider the long-term societal implications of psychometric targeting and data-driven misinformation (Townsend and Wallace, 2016).

45. Final Reflection

The digital landscape is poised for continued expansion, with the ongoing advancements in artificial intelligence, large language models, and machine learning signalling a trajectory of relentless technological development. These advancements aim to improve societal well-being, foster human progress, and support more sustainable ways of living. Nonetheless, as technological innovation progresses, the conflict between innovation and ethical responsibility becomes increasingly prominent. Recent research provides evidence of both opportunities and risks linked to psychometric profiling in micro-targeted advertising, many involving misleading or false information that could sway behaviour. This highlights the need for collective responsibility to ensure data science practices adhere to core principles of democracy, transparency, and the well-being of users.

This research faced several challenges, notably due to CA's involvement in a prominent scandal. Facebook's mishandling of third-party applications significantly contributed to these issues. Moreover, the CA scandal highlights the tendency of large institutions, such as Facebook, to treat user data as commodities, primarily for targeted advertising and monetisation, often ignoring user interests. The theoretical analysis was particularly demanding, as initial assumptions about data accessibility proved incorrect; it later became apparent that the ICO withheld all relevant information. This withholding can be justified because such methodologies, if widely available, could be exploited by malicious actors to imitate harmful practices.

Reference:

- Adeniran, I.A., Efunniyi, C.P., Osundare, O.S. and Abhulimen, A.O. (2024). The role of data science in transforming business operations: Case studies from enterprises. *Computer Science & IT Research Journal*, [online] 5(8), pp.2026–2039.
doi:<https://doi.org/10.51594/csitrj.v5i8.1490> [Accessed 11 Aug. 2025].
- Afsharian, M. (2025). Data science essentials in business administration: A multidisciplinary perspective. *Decision analytics journal*, [online] pp.100442–100442.
doi:<https://doi.org/10.1016/j.dajour.2024.100442> [Accessed 11 Aug. 2025].
- Akbar, M.R., Huma, T. and Rizvee, S.M. (2025). Traditional Television Advertising Versus Digital Advertising: A Comparative Study on Audience Attention, Engagement, and Persuasiveness in the Modern Era: <https://doi.org/10.5281/zenodo.16965295>. *Dialogue Social Science Review (DSSR)*, [online] 3(8), pp.599–618.
doi:<https://doi.org/10.5281/zenodo.16965295> [Accessed 13 Oct. 2025].
- Albright, J. (2018). *The Graph API: Key Points in the Facebook and Cambridge Analytica Debacle*. [online] Medium. Available at: <https://medium.com/tow-center/the-graph-api-key-points-in-the-facebook-and-cambridge-analytica-debacle-b69fe692d747> [Accessed 4 Sep. 2025].
- Alsaleh, A. (2024). The impact of technological advancement on culture and society. *Scientific Reports*, [online] 14(1). doi:<https://doi.org/10.1038/s41598-024-83995-z> [Accessed 8 Aug. 2025].
- Amazon Mechanical Turk (2018). *Amazon Mechanical Turk*. [online] Mturk.com. Available at: <https://www.mturk.com/> [Accessed 7 Sep. 2025].
- Anderson, C., Sharps, D.L., Soto, C.J. and John, O.P. (2020). People with disagreeable personalities (selfish, combative, and manipulative) do not have an advantage in pursuing power at work. *Proceedings of the National Academy of Sciences*, [online] 117(37), pp.22780–22786. doi:<https://doi.org/10.1073/pnas.2005088117> [Accessed 3 Oct. 2025].
- Arvai, K. (2020). *K-Means Clustering in Python: A Practical Guide*. [online] realpython.com. Available at: <https://realpython.com/k-means-clustering-python/> [Accessed 20 Sep. 2025].
- Bakir, V. (2020). Psychological Operations in Digital Political Campaigns: Assessing Cambridge Analytica's Psychographic Profiling and Targeting. *Frontiers in Communication*, [online] 5. doi:<https://doi.org/10.3389/fcomm.2020.00067> [Accessed 9 Oct. 2025].
- Ballhaus, R. (2017). Trump-Linked Company Reached Out to WikiLeaks on Hacked Emails. *Wall Street Journal*. [online] 25 Oct. Available at: <https://www.wsj.com/articles/wikileaks-assange-says-he-rejected-overture-from-trump-linked-group-1508961298> [Accessed 3 Sep. 2025].

Barrett, B. (2018). *How to Check If Cambridge Analytica Could Access Your Facebook Data*. [online] WIRED. Available at: <https://www.wired.com/story/did-cambridge-analytica-access-your-facebook-data/> [Accessed 4 Sep. 2025].

Barth, S. and de Jong, M.D.T. (2017). The Privacy Paradox – Investigating Discrepancies between Expressed Privacy Concerns and Actual Online Behavior – a Systematic Literature Review. *Telematics and Informatics*, [online] 34(7), pp.1038–1058. doi:<https://doi.org/10.1016/j.tele.2017.04.013> [Accessed 10 Sep. 2025].

BBC (2018a). ‘Cambridge Analytica planted fake news’. *BBC News*. [online] 20 Mar. Available at: <https://www.bbc.co.uk/news/av/world-43472347> [Accessed 19 Aug. 2025].

BBC (2018b). Cambridge Analytica offices searched over data storage. *BBC News*. [online] 23 Mar. Available at: <https://www.bbc.co.uk/news/uk-43522775> [Accessed 22 Sep. 2025].

BBC Bitesize (2024). *Timeline of how online misinformation fuelled UK riots - BBC Bitesize*. [online] BBC Bitesize. Available at: <https://www.bbc.co.uk/bitesize/articles/zshjs82> [Accessed 9 Oct. 2025].

BBC News (2018a). Cambridge Analytica: Facebook boss summoned over data claims. *BBC News*. [online] 20 Mar. Available at: <https://www.bbc.com/news/uk-43474760> [Accessed 14 Oct. 2025].

BBC News (2018b). The global reach of Cambridge Analytica. *BBC News*. [online] 22 Mar. Available at: <https://www.bbc.com/news/world-43476762> [Accessed 14 Oct. 2025].

Benesch, S. (2021). *Nobody Can See Into Facebook*. [online] Available at: http://www.businessforum.com/Atlantic_10-30-2021.pdf [Accessed 11 Aug. 2025].

Boyce, R., Whelan, R. and Hashemi, R. (2025). *Here's how data awareness can tackle cybercrime's complexity*. [online] World Economic Forum. Available at: <https://www.weforum.org/stories/2025/01/cybercrime-data-cybersecurity/> [Accessed 23 Jul. 2025].

British Psychological Society (2021). Code of ethics and conduct. *British Psychological Society*, [online] pp.1–10. doi:<https://doi.org/10.53841/bpsrep.2021.inf94> [Accessed 9 Sep. 2025].

Byte Myke (2021). *SQLite beginner crash course in Visual Studio Code - 2022*. [online] YouTube. Available at: https://www.youtube.com/watch?v=IBgWKTaG_Bs [Accessed 19 Oct. 2025].

Cadwalladr, C. (2019). *Cambridge Analytica a year on: ‘a lesson in institutional failure’*. [online] the Guardian. Available at: <https://www.theguardian.com/uk-news/2019/mar/17/cambridge-analytica-year-on-lesson-in-institutional-failure-christopher-wylie> [Accessed 3 Sep. 2025].

Cadwalladr, C. and Graham-Harrison, E. (2018). *Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach*. [online] The Guardian. Available at: <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election> [Accessed 3 Aug. 2025].

Cambridge Dictionary (2019). *DATA | meaning in the Cambridge English Dictionary*. [online] Cambridge.org. Available at: <https://dictionary.cambridge.org/dictionary/english/data> [Accessed 28 Jul. 2025].

Cawthra, J., Ekstrom, M., Lusty, L., Sexton, J., Sweetnam, J. and Townsend, A. (2020). *Data Integrity: Detecting and Responding to Ransomware and Other Destructive Events*. [online] www.nccoe.nist.gov. Available at: <https://www.nccoe.nist.gov/publication/1800-26/VolA/index.html> [Accessed 7 Sep. 2025].

Chacón, A., Borda-Mas, M., Rivera, F., Pérez-Chacón, M. and María Luisa Avargues-Navarro (2024). Aesthetic sensitivity: relationship with openness to experience and agreeableness, health-related quality of life and adaptive coping strategies in people with high sensory processing sensitivity. *Frontiers in Psychology*, [online] 14. doi:<https://doi.org/10.3389/fpsyg.2023.1276124> [Accessed 3 Oct. 2025].

Chang, A. (2018). *The Facebook and Cambridge Analytica scandal, explained with a simple diagram*. [online] Vox. Available at: <https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram> [Accessed 27 Aug. 2025].

Chauliac, M., Willems, J., Gijbels, D. and Donche, V. (2023). The prevalence of careless response behaviour and its consequences on data quality in self-report questionnaires on student learning. *Frontiers in Education*, [online] 8. doi:<https://doi.org/10.3389/feduc.2023.1197324> [Accessed 4 Oct. 2025].

Chinthala, L.K. (2023). Next-Gen Marketing: Trends in Influencer Marketing, Data-Driven Campaigns, and Social Media Evolution. *International Journal of Scientific Research & Engineering Trends* , [online] 9(2). Available at: https://www.researchgate.net/profile/Lakshmi-Chinthala/publication/391718585_Next-Gen_Marketing_Trends_in_Influencer_Marketing_Data-Driven_Campaigns_and_Social_Media_Evolution/links/682635de6b5a287c3041e8c4/Next-Gen-Marketing-Trends-in-Influencer-Marketing-Data-Driven-Campaigns-and-Social-Media-Evolution.pdf [Accessed 31 Aug. 2025].

Choi, H., Park, J. and Jung, Y. (2018). The role of privacy fatigue in online privacy behavior. *Computers in Human Behavior*, [online] 81(81), pp.42–51. doi:<https://doi.org/10.1016/j.chb.2017.12.001> [Accessed 10 Sep. 2025].

Codecademy (2019). *What is an IDE? Understanding Integrated Development Environments*. [online] Codecademy. Available at: <https://www.codecademy.com/article/what-is-ide> [Accessed 29 Sep. 2025].

Confessore, N. (2018). Cambridge Analytica and Facebook: the Scandal and the Fallout so Far. *The New York Times*. [online] 4 Apr. Available at: <https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html> [Accessed 14 Oct. 2025].

Constine, J. (2015). *Facebook Is Shutting Down Its API For Giving Your Friends' Data To Apps*. [online] TechCrunch. Available at: <https://techcrunch.com/2015/04/28/facebook-api-shut-down/> [Accessed 8 Sep. 2025].

CyBOK (n.d.). *CyBOK – The Cyber Security Body of Knowledge at a glance*. [online] www.cybok.org. Available at: <https://www.cybok.org/ataglance/> [Accessed 31 Aug. 2025].

Dam, V.H., Hjordt, L.V., Cunha-Bang, S., Sestoft, D., Knudsen, G.M. and Stenbæk, D.S. (2021). Trait aggression is associated with five-factor personality traits in males. *Brain and Behavior*, [online] 11(7). doi:<https://doi.org/10.1002/brb3.2175> [Accessed 3 Oct. 2025].

Data Science Discovery (n.d.). *Descriptive Statistics*. [online] Data Science Discovery. Available at: <https://discovery.cs.illinois.edu/learn/Exploratory-Data-Analysis/Descriptive-Statistics/> [Accessed 29 Sep. 2025].

Davies, R. and Rushe, D. (2019). *Facebook to pay \$5bn fine as regulator settles Cambridge Analytica complaint*. [online] The Guardian. Available at: <https://www.theguardian.com/technology/2019/jul/24/facebook-to-pay-5bn-fine-as-regulator-files-cambridge-analytica-complaint> [Accessed 15 Sep. 2025].

Deuker, A. (2010). Addressing the Privacy Paradox by Expanded Privacy Awareness – The Example of Context-Aware Services. *IFIP Advances in Information and Communication Technology*, [online] pp.275–283. doi:https://doi.org/10.1007/978-3-642-14282-6_23 [Accessed 10 Sep. 2025].

Dhiman, B. (2023). Ethical Issues and Challenges in Social media: A Current Scenario. *Global Media Journal*, [online] 21(62), pp.1–5. doi:<https://doi.org/10.36648/1550-7521.21.62.368> [Accessed 10 Sep. 2025].

Diener, E., Lucas, R.E. and Cummings, J.A. (2019). *16.1 Personality Traits*. [online] Saskoer.ca. Available at: <https://www.saskoer.ca/introductiontopsychology/chapter/personality-traits/> [Accessed 4 Oct. 2025].

Dixon, S.J. (2024). *Facebook Users Reach by Device 2019 | Statista*. [online] Statista. Available at: <https://www.statista.com/statistics/377808/distribution-of-facebook-users-by-device/> [Accessed 10 Aug. 2025].

Dixon, S.J. (2025a). *Distribution of Facebook users worldwide as of April 2024, by age and gender*. [online] Statista. Available at: <https://www.statista.com/statistics/376128/facebook-global-user-age-distribution/> [Accessed 10 Aug. 2025].

Dixon, S.J. (2025b). *Global Facebook users by region 2020* | Statista. [online] Statista. Available at: <https://www.statista.com/statistics/273506/comparison-of-unique-visitors-to-facebook-sorted-by-region/#statisticContainer> [Accessed 10 Aug. 2025].

Dixon, S.J. (2025c). *Most Popular Social Networks Worldwide as of February 2025, by Number of Monthly Active Users*. [online] Statista. Available at: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/> [Accessed 4 Aug. 2025].

Dowling, M.-E. (2022). Cyber information operations: Cambridge Analytica's challenge to democratic legitimacy. *Journal of Cyber Policy*, [online] 7(2), pp.1–19. doi:<https://doi.org/10.1080/23738871.2022.2081089> [Accessed 14 Sep. 2025].

EDHEC Online (2025). *Applying Data Ethics: A Practical Guide for Responsible Data Use*. [online] EDHEC Online. Available at: <https://online.edhec.edu/en/blog/applying-data-ethics-a-practical-guide-for-responsible-data-use/> [Accessed 12 Aug. 2025].

Epstein, D. and Medzini, R. (2022). Conversations with fellow leaders: Privacy framing in congressional hearings after Cambridge Analytica. *Telecommunications Policy*, [online] 46(10), p.102427. doi:<https://doi.org/10.1016/j.telpol.2022.102427> [Accessed 10 Sep. 2025].

Federal Trade Commission (2019). *FTC Imposes \$5 Billion Penalty and Sweeping New Privacy Restrictions on Facebook*. [online] Federal Trade Commission. Available at: <https://www.ftc.gov/news-events/news/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions-facebook> [Accessed 18 Sep. 2025].

Fernando, J. (2021). *Cambridge Analytica*. [online] Investopedia. Available at: <https://www.investopedia.com/terms/c/cambridge-analytica.asp> [Accessed 2 Sep. 2025].

Filmology (2025). *TV Ads vs Digital Ads in 2025: Which One Truly Wins?* [online] Filmology Productions. Available at: <https://filmology.com.sa/tv-ads-vs-digital-ads/> [Accessed 13 Oct. 2025].

Floridi, L. (2021). *Data*. In William A. Darity, *International Encyclopedia of the Social Sciences*. [online] Philarchive.org. Available at: <https://philarchive.org/rec/FLOD-2> [Accessed 28 Jul. 2025].

free-news-api (2024). *GitHub - free-news-api/news-api: Top Free News API Comparison*. [online] GitHub. Available at: <https://github.com/free-news-api/news-api> [Accessed 18 Oct. 2025].

Fung, B. (2024). *UK riots show how social media can fuel real-life harm. It's only getting worse.* [online] CNN. Available at: <https://edition.cnn.com/2024/08/09/tech/uk-protests-social-media> [Accessed 9 Oct. 2025].

- Galan, S. (2025). *Global population from 2000 to 2023, by gender*. [online] Statista. Available at: <https://www.statista.com/statistics/1328107/global-population-gender/> [Accessed 10 Aug. 2025].
- GeeksforGeeks (2021). *Pandas Scatter Plot – DataFrame.plot.scatter()*. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/python/pandas-scatter-plot-dataframe-plot-scatter/> [Accessed 23 Sep. 2025].
- GeeksforGeeks (2024). *Introduction to Python Pytesseract Package*. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/python/introduction-to-python-ptesseract-package/> [Accessed 18 Oct. 2025].
- González-Pizarro, F., Figueroa, A., López, C. and Aragon, C. (2022). Regional Differences in Information Privacy Concerns After the Facebook-Cambridge Analytica Data Scandal. *Computer Supported Cooperative Work (CSCW)*, [online] 31, pp.33–77. doi:<https://doi.org/10.1007/s10606-021-09422-3> [Accessed 22 Aug. 2025].
- Google Colab (2019). *Google Colaboratory*. [online] Google.com. Available at: <https://colab.research.google.com/> [Accessed 20 Sep. 2025].
- GOV.uk (n.d.). *CAMBRIDGE ANALYTICA(UK) LIMITED - Overview (free company information from Companies House)*. [online] find-and-update.company-information.service.gov.uk. Available at: <https://find-and-update.company-information.service.gov.uk/company/09375920> [Accessed 3 Aug. 2025].
- Grady, C. (2015). Institutional Review Boards. *Chest*, [online] 148(5), pp.1148–1155. doi:<https://doi.org/10.1378/chest.15-0706> [Accessed 14 Sep. 2025].
- GWI (2025). *The 2025 social media report | GWI*. [online] Gwi.com. Available at: <https://www.gwi.com/reports/social-media-trends/explore?submissionGuid=3d4f02fd-2b6d-4ffe-9950-72df504a2b51> [Accessed 10 Aug. 2025].
- Hartmans, A. (2018). *It's impossible to know exactly what data Cambridge Analytica scraped from Facebook — but here's the kind of information apps could access in 2014*. [online] Business Insider. Available at: <https://www.businessinsider.com/what-data-did-cambridge-analytica-have-access-to-from-facebook-2018-3> [Accessed 4 Sep. 2025].
- Hayes, A. (2024). *Descriptive statistics: Definition, overview, types, example*. [online] Investopedia. Available at: https://www.investopedia.com/terms/d/descriptive_statistics.asp [Accessed 29 Sep. 2025].
- Hewood, J. (2018). Pseudo-public political speech: Democratic implications of the Cambridge Analytica scandal. *Information Polity*, [online] 23(4), pp.429–434. doi:<https://doi.org/10.3233/IP-180009> [Accessed 14 Sep. 2025].

Hern, A. (2019). *Facebook agrees to pay fine over Cambridge Analytica scandal*. [online] the Guardian. Available at: <https://www.theguardian.com/technology/2019/oct/30/facebook-agrees-to-pay-fine-over-cambridge-analytica-scandal> [Accessed 15 Sep. 2025].

Hern, A. and Cadwalladr, C. (2018). *Revealed: Aleksandr Kogan collected Facebook users' direct messages*. [online] the Guardian. Available at: <https://www.theguardian.com/uk-news/2018/apr/13/revealed-aleksandr-kogan-collected-facebook-users-direct-messages> [Accessed 5 Sep. 2025].

HIIG (2018). *The ethics of big data, Facebook & Cambridge Analytica | Digital Society Blog*. [online] The Alexander von Humboldt Institute for Internet and Society (HIIG). Available at: <https://www.hiig.de/en/ethics-big-data-facebook-cambridge-analytica/> [Accessed 24 Aug. 2025].

Hill, P.L. and Edmonds, G.W. (2017). Personality development in adolescence. *Personality Development Across the Lifespan*, [online] pp.25–38. doi:<https://doi.org/10.1016/b978-0-12-804674-6.00003-x> [Accessed 28 Sep. 2025].

Hinds, J., Williams, E.J. and Joinson, A.N. (2020). ‘It wouldn’t happen to me’: Privacy concerns and perspectives following the Cambridge Analytica scandal. *International Journal of Human-Computer Studies*, [online] 143(2020), p.102498. doi:<https://doi.org/10.1016/j.ijhcs.2020.102498> [Accessed 9 Sep. 2025].

History.com Editors (2018). *The 2016 U.S. Presidential Election*. [online] HISTORY. Available at: <https://www.history.com/articles/us-presidential-election-2016> [Accessed 3 Sep. 2025].

Hitlin, P. and Rainie, L. (2011). *Facebook Algorithms and Personal Data*. [online] Pew Research Center, pp.1–23. Available at: <https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/> [Accessed 10 Aug. 2025].

Hu, M. (2020). Cambridge Analytica’s black box. *Big Data & Society*, [online] 7(2), pp.1–6. doi:<https://doi.org/10.1177/2053951720938091> [Accessed 22 Aug. 2025].

Hunt, D. and Messinger, P.R. (2018). Cambridge Analytica, influencing elections and the INFORMS Ethics Guidelines. [online] doi:<https://doi.org/10.1287/orms.2018.05.10> [Accessed 2 Sep. 2025].

ICO (2023). *A Guide to the Data Protection Principles*. [online] Information Commissioner's Office. Available at: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/> [Accessed 23 Sep. 2025].

Ingram, D. (2018). Factbox: Who is Cambridge Analytica and what did it do? *Reuters*. [online] 20 Mar. Available at: <https://www.reuters.com/article/technology/factbox-who-is-cambridge-analytica-and-what-did-it-do-idUSKBN1GW07F/> [Accessed 1 Sep. 2025].

Isaak, J. and Hanna, M.J. (2018). User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer*, [online] 51(8), pp.56–59.
doi:<https://doi.org/10.1109/MC.2018.3191268> [Accessed 9 Sep. 2025].

Jackson, J.J., Wood, D., Bogg, T., Walton, K.E., Harms, P.D. and Roberts, B.W. (2010). What do conscientious people do? Development and validation of the Behavioral Indicators of Conscientiousness (BIC). *Journal of Research in Personality*, [online] 44(4), pp.501–511.
doi:<https://doi.org/10.1016/j.jrp.2010.06.005> [Accessed 3 Oct. 2025].

Joseph, J., Sivaraman, J., Periyasamy, R. and Simi, V.R. (2017). An objective method to identify optimum clip-limit and histogram specification of contrast limited adaptive histogram equalization for MR images. *Biocybernetics and Biomedical Engineering*, 37(3), pp.489–497. doi:<https://doi.org/10.1016/j.bbe.2016.11.006> [Accessed 19 Oct. 2025].

Kati, E. (2022). *What is a Shell Company?* | LegalVision UK. [online] legalvision.co.uk. Available at: <https://legalvision.co.uk/corporations/what-is-a-shell-company/> [Accessed 1 Sep. 2025].

Kekäläinen, T., Terracciano, A., Sipilä, S. and Kokko, K. (2020). Personality traits and physical functioning: a cross-sectional multimethod facet-level analysis. *European Review of Aging and Physical Activity*, [online] 17(1). doi:<https://doi.org/10.1186/s11556-020-00251-9> [Accessed 29 Sep. 2025].

Kenber, B. (2018). *Cambridge Analytica data inquiry to continue despite bankruptcy*. [online] Thetimes.com. Available at: <https://www.thetimes.com/business-money/technology/article/cambridge-analytica-closes-after-data-harvesting-scandal-jpshm3sm6> [Accessed 5 Sep. 2025].

Kenton, W. (2019). *Shell Corporation*. [online] Investopedia. Available at: <https://www.investopedia.com/terms/s/shellcorporation.asp> [Accessed 1 Sep. 2025].

Kiderlin, S. (2024). *Online disinformation sparked a wave of far-right violence in the UK — here's how*. [online] CNBC. Available at: <https://www.cnbc.com/2024/08/09/online-disinformation-sparked-a-wave-of-far-right-violence-in-the-uk.html> [Accessed 9 Oct. 2025].

Kite (2000). *Sqlite 3 Python Tutorial in 5 minutes - Creating Database, Tables and Querying [2020]*. [online] www.youtube.com. Available at: <https://www.youtube.com/watch?v=girsuXz0yA8> [Accessed 18 Oct. 2025].

Krainyk, Y., Razzhyvin, A., Bondarenko, O. and Simakova, I. (2019). Internet-of-Things Device Set Configuration for Connection to Wireless Local Area Network. *Computer Modeling and Intelligent Systems*, 2353, pp.885–896.
doi:<https://doi.org/10.32782/cmis/2353-70> [Accessed 8 Aug. 2025].

Kumar, N. (2025). *Facebook Users Statistics 2024 (Worldwide Data)*. [online] demandsage. Available at: <https://www.demandsage.com/facebook-statistics/> [Accessed 10 Aug. 2025].

Lal, B., Dwivedi, Y.K. and Haag, M. (2021). Working from Home during Covid-19: Doing and Managing Technology-enabled Social Interaction with Colleagues at a Distance. *Information Systems Frontiers*, [online] 25, pp.1333–1350. doi:<https://doi.org/10.1007/s10796-021-10182-0> [Accessed 8 Aug. 2025].

Lauer, D. (2021). Facebook's ethical failures are not accidental; they are part of the business model. *AI and Ethics*, [online] 1(1), pp.395–403. doi:<https://doi.org/10.1007/s43681-021-00068-x> [Accessed 8 Aug. 2025].

Lee, M. (2022). *pytesseract: Python-tesseract is a python wrapper for Google's Tesseract-OCR*. [online] PyPI. Available at: <https://pypi.org/project/pytesseract/> [Accessed 18 Oct. 2025].

Legislation.Gov.UK (2023). *Online Safety Act 2023*. [online] Legislation.gov.uk. Available at: <https://www.legislation.gov.uk/ukpga/2023/50> [Accessed 9 Oct. 2025].

Leonelli, S. (2016). Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, [online] 374(2083), p.20160122. doi:<https://doi.org/10.1098/rsta.2016.0122> [Accessed 14 Oct. 2025].

Lewis, P. and Wong, J.C. (2018). Facebook employs psychologist whose firm sold data to Cambridge Analytica. *The Guardian*. [online] 18 Mar. Available at: <https://www.theguardian.com/news/2018/mar/18/facebook-cambridge-analytica-joseph-chancellor-gsr> [Accessed 5 Sep. 2025].

Lim, A. (2025). Big Five Personality Traits: The 5-Factor Model Of Personality. *Simply Psychology*. [online] Available at: <https://www.simplypsychology.org/big-five-personality.html> [Accessed 1 Oct. 2025].

Lindridge, J. (2017). Principlism: When Values Conflict. *Journal of Paramedic Practice*, [online] 9(4), pp.158–163. doi:<https://doi.org/10.12968/jpar.2017.9.4.158> [Accessed 13 Oct. 2025].

Lusha, E. (2023). The Impact of Technological Advancements on Society: Examining the Possibility of a Brave New World. *Economicus*, [online] 22(2), p.57. doi:<https://doi.org/10.58944/iblr9210> [Accessed 8 Aug. 2025].

Lynwood, W. (2025). *Using data from the internet and social media in research: ethics & consent*. [online] Available at: <https://info.lse.ac.uk/staff/divisions/research-and-innovation/research/Assets/Documents/PDF/ethics-Using-internet-and-Social-media-data.pdf> [Accessed 25 Aug. 2025].

- Manokha, I. (2025). *Surveillance: The DNA of Platform Capital—The Case of Cambridge Analytica Put into Perspective*. [online] Oclc.org. Available at: <https://muse-jhu-edu.uniessexlib.idm.oclc.org/article/707015> [Accessed 11 Sep. 2025].
- Markham, A. and Buchanan, E. (2012). *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)*. [online] Available at: <https://aoir.org/reports/ethics2.pdf> [Accessed 8 Sep. 2025].
- Marotta, A. and Madnick, S. (2025). Analyzing and Categorizing Emerging Cybersecurity Regulations. *ACM Computing Surveys*, [online] 58(2). doi:<https://doi.org/10.1145/3757318> [Accessed 14 Oct. 2025].
- Matplotlib (2025). *scatter(x, y) — Matplotlib 3.7.1 documentation*. [online] matplotlib.org. Available at: https://matplotlib.org/stable/plot_types/basic/scatter_plot.html#sphx-glr-plot-types-basic-scatter-plot-py [Accessed 25 Sep. 2025].
- mattlisiv (2018). *GitHub - mattlisiv/newsapi-python: A Python Client for News API*. [online] GitHub. Available at: <https://github.com/mattlisiv/newsapi-python> [Accessed 19 Oct. 2025].
- McCrae, R.R. and John, O.P. (1992). An Introduction to the five-factor Model and Its Applications. *Journal of Personality*, [online] 60(2), pp.175–215. doi:<https://doi.org/10.1111/j.1467-6494.1992.tb00970.x> [Accessed 21 Sep. 2025].
- Meta (2018). *We're Making Our Terms and Data Policy Clearer, Without New Rights to Use Your Data on Facebook*. [online] About Facebook. Available at: <https://about.fb.com/news/2018/04/terms-and-data-policy/> [Accessed 4 Aug. 2025].
- Meta (2022). *Facebook Data policy*. [online] Facebook. Available at: <https://www.facebook.com/about/privacy/update/printable> [Accessed 31 Aug. 2025].
- Meta (2025). *Permissions Reference for Meta Technologies APIs*. [online] Facebook.com. Available at: <https://developers.facebook.com/docs/permissions/> [Accessed 8 Sep. 2025].
- Microsoft Excel (2024). *Microsoft Excel, Spreadsheet Software*. [online] www.microsoft.com. Available at: <https://www.microsoft.com/en-us/microsoft-365/excel> [Accessed 29 Sep. 2025].
- Mohamed, E. (2024). *Southport stabbing: What led to the spread of disinformation?* [online] Al Jazeera. Available at: <https://www.aljazeera.com/news/2024/8/2/southport-stabbing-what-led-to-the-spread-of-disinformation> [Accessed 9 Oct. 2025].
- Nasello, J., Triffaux, J.-M. and Hansenne, M. (2023). Individual differences and personality traits across situations. *Current Issues in Personality Psychology*, [online] 12(2). doi:<https://doi.org/10.5114/cipp/159942> [Accessed 2 Oct. 2025].
- Netflix (2019). *The Great Hack | Netflix Official Site*. [online] www.netflix.com. Available at: <https://www.netflix.com/gb/title/80117542> [Accessed 8 Sep. 2025].

Newsapi.org (2019). *News API - A JSON API for live news and blog articles*. [online] Newsapi.org. Available at: <https://newsapi.org/> [Accessed 19 Oct. 2025].

NewsData (n.d.). *NewsData - News API to Search & Collect Worldwide News*. [online] Newsdata. Available at: <https://newsdata.io/> [Accessed 19 Oct. 2025].

Ng, D.X., Lin, P.K.F., Marsh, N.V., Chan, K.Q. and Ramsay, J.E. (2021). Associations Between Openness Facets, Prejudice, and Tolerance: A Scoping Review With Meta-Analysis. *Frontiers in Psychology*, 12, p.<https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.707652/full>. doi:<https://doi.org/10.3389/fpsyg.2021.707652> [Accessed 3 Oct. 2025].

NumPy (2022). *NumPy Documentation*. [online] numpy.org. Available at: <https://numpy.org/doc/> [Accessed 19 Oct. 2025].

Nyabola, N. (2019). *The spectre of Cambridge Analytica still haunts African elections*. [online] Al Jazeera. Available at: <https://www.aljazeera.com/opinions/2019/2/15/the-spectre-of-cambridge-analytica-still-haunts-african-elections> [Accessed 14 Oct. 2025].

O'Hagan, E.M. (2018). *No one can pretend Facebook is just harmless fun any more*. [online] the Guardian. Available at: <https://www.theguardian.com/commentisfree/2018/mar/18/facebook-extremist-content-user-data> [Accessed 13 Aug. 2025].

OECD (2023). *Review of the OECD recommendation on Cross-Border Co-Operation in the Enforcement of Laws Protecting Privacy*. [online] Available at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2023/09/review-of-the-oecd-recommendation-on-cross-border-co-operation-in-the-enforcement-of-laws-protecting-privacy_fc556c18/67774f69-en.pdf [Accessed 22 Aug. 2025].

Opencv.org (n.d.). *OpenCV documentation index*. [online] docs.opencv.org. Available at: <https://docs.opencv.org/> [Accessed 19 Oct. 2025].

Opoku, O.G., Adamu, A. and Daniel, O. (2023). Relation between students' personality traits and their preferred teaching methods: Students at the university of Ghana and the Huzhou Normal University. *Heliyon*, [online] 9(1), p.e13011. doi:<https://doi.org/10.1016/j.heliyon.2023.e13011> [Accessed 3 Oct. 2025].

Pandas (2025). *User Guide — pandas 1.0.4 documentation*. [online] pandas.pydata.org. Available at: https://pandas.pydata.org/docs/user_guide/index.html#user-guide [Accessed 22 Sep. 2025].

Pillow.readthedocs.io (n.d.). *Pillow — Pillow (PIL Fork) 7.2.0 documentation*. [online] pillow.readthedocs.io. Available at: <https://pillow.readthedocs.io/> [Accessed 19 Oct. 2025].

Prichard, E.C. (2021). Is the Use of Personality Based Psychometrics by Cambridge Analytical Psychological Science's 'Nuclear Bomb' Moment?. *Frontiers in Psychology*, [online] 12. doi:<https://doi.org/10.3389/fpsyg.2021.581448> [Accessed 14 Sep. 2025].

PyPi (2019). *opencv-python*. [online] PyPI. Available at: <https://pypi.org/project/opencv-python/> [Accessed 18 Oct. 2025].

Python Software Foundation (2024). *sqlite3 — DB-API 2.0 interface for SQLite databases — Python 3.8.2 documentation*. [online] docs.python.org. Available at: <https://docs.python.org/3/library/sqlite3.html> [Accessed 19 Oct. 2025].

Python.org (n.d.). *difflib — Helpers for computing deltas*. [online] Python documentation. Available at: <https://docs.python.org/3/library/difflib.html#difflib.SequenceMatcher> [Accessed 19 Oct. 2025].

Renotte, N. (2021). *How to Install OpenCV for Python // OpenCV for Beginners*. [online] YouTube. Available at: <https://www.youtube.com/watch?v=M6jukmppMqU> [Accessed 18 Oct. 2025].

Requests.readthedocs.io (n.d.). *Requests: HTTP for Humans™ — Requests 2.31.0 documentation*. [online] requests.readthedocs.io. Available at: <https://requests.readthedocs.io/> [Accessed 19 Oct. 2025].

Resnick, B. (2018). *Cambridge Analytica's 'psychographic microtargeting': what's bullshit and what's legit*. [online] Vox. Available at: <https://www.vox.com/science-and-health/2018/3/23/17152564/cambridge-analytica-psychographic-microtargeting-what> [Accessed 9 Oct. 2025].

Romano, A. (2018). *The Facebook data breach wasn't a hack. It was a wake-up call*. [online] Vox. Available at: <https://www.vox.com/2018/3/20/17138756/facebook-data-breach-cambridge-analytica-explained> [Accessed 3 Sep. 2025].

Rosenberg, D. (n.d.). *Data before the Fact*. [online] Available at: https://projects.iq.harvard.edu/sites/projects.iq.harvard.edu/files/eswg/files/rosenborg_-_rawdata.pdf [Accessed 28 Jul. 2025].

Rosenberg, M., Confessore, N. and Cadwalladr, C. (2018). How Trump Consultants Exploited the Facebook Data of Millions. *The New York Times*. [online] 17 Mar. Available at: <https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html> [Accessed 10 Sep. 2025].

SBS News (2018). *'Likely' Australians caught up in Cambridge Analytica data scandal*. [online] SBS News. Available at: <https://www.sbs.com.au/news/article/likely-australians-caught-up-in-cambridge-analytica-data-scandal/bnh3h7olz> [Accessed 14 Oct. 2025].

Schneble, C.O., Elger, B.S. and Shaw, D. (2018). The Cambridge Analytica Affair and Internet-mediated Research. *EMBO reports*, [online] 19(8). doi:<https://doi.org/10.15252/embr.201846579> [Accessed 22 Aug. 2025].

Schulz, J. (2013). Geometric optics and strategies for subsea imaging. [online] doi:<https://doi.org/10.1533/9780857093523.3.243> [Accessed 19 Oct. 2025].

Scikit-learn (2019). *sklearn.cluster.KMeans — scikit-learn 0.21.3 documentation*. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html> [Accessed 21 Sep. 2025].

Seng, S., Al-Ameen, M.N. and Wright, M. (2021). A look into user privacy and third-party applications in Facebook. [online] 29(2), pp.283–313. doi:<https://doi.org/10.1108/ics-08-2019-0108> [Accessed 31 Aug. 2025].

Shah, M. (2024). *Fanning the Flames: Online Misinformation and Far-Right Violence in the UK - GNET*. [online] GNET. Available at: <https://gnet-research.org/2024/08/28/fanning-the-flames-online-misinformation-and-far-right-violence-in-the-uk/> [Accessed 9 Oct. 2025].

Sheikh, M. (2025). *Social Media Demographics to Inform Your 2025 Strategy*. [online] Sprout Social. Available at: <https://sproutsocial.com/insights/new-social-media-demographics/> [Accessed 10 Aug. 2025].

Siegelman, W. (2018). *Cambridge Analytica is dead – but its obscure network is alive and well*. [online] the Guardian. Available at: <https://www.theguardian.com/uk-news/2018/may/05/cambridge-analytica-scl-group-new-companies-names> [Accessed 2 Sep. 2025].

Socratica (2023). *SQLite in Python || Python Tutorial || Learn Python Programming*. [online] www.youtube.com. Available at: <https://www.youtube.com/watch?v=c8yHTlrs9EA> [Accessed 18 Oct. 2025].

Soto, C.J. and John, O.P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, [online] 113(1), pp.117–143. doi:<https://doi.org/10.1037/pspp0000096> [Accessed 28 Sep. 2025].

SQLite (2014). *Datatypes In SQLite Version 3*. [online] Sqlite.org. Available at: <https://www.sqlite.org/datatype3.html> [Accessed 19 Oct. 2025].

Stieglitz, S., Mirbabaie, M., Ross, B. and Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, [online] 39(39), pp.156–168. Available at: <https://www.sciencedirect.com/science/article/pii/S0268401217308526> [Accessed 10 Aug. 2025].

Sutton, J. (2025). *Big Five Personality Traits: The OCEAN Model Explained [2019 Upd.]*. [online] PositivePsychology.com. Available at: <https://positivepsychology.com/big-five-personality-theory/> [Accessed 2 Oct. 2025].

Tarran, B. (2018). What can we learn from the Facebook-Cambridge Analytica scandal? *Significance*, [online] 15(3), pp.4–5. doi:<https://doi.org/10.1111/j.1740-9713.2018.01139.x> [Accessed 11 Sep. 2025].

Tech Policy Press (2024). *In the Matter of Facebook, Inc.* | TechPolicy.Press. [online] Tech Policy Press. Available at: <https://www.techpolicy.press/tracker/in-the-matter-of-facebook-inc/> [Accessed 18 Sep. 2025].

Tesseract-ocr.github (2020). *Tesseract User Manual*. [online] tessdoc. Available at: <https://tesseract-ocr.github.io/tessdoc/> [Accessed 19 Oct. 2025].

The Guardian (2018). *Cambridge Analytica whistleblower: ‘We spent \$1m harvesting millions of Facebook profiles’*. YouTube. Available at: <https://www.youtube.com/watch?v=FXdYSQ6nu-M> [Accessed 3 Aug. 2025].

The New York Times (2017). Presidential Election Results: Donald J. Trump Wins. *The New York Times*. [online] 9 Aug. Available at: <https://www.nytimes.com/elections/2016/results/president> [Accessed 3 Sep. 2025].

The News API (2025). *Free live and top story JSON news API* | The News API. [online] Thenewsapi.com. Available at: <https://www.thenewsapi.com/> [Accessed 19 Oct. 2025].

Townsend, L. and Wallace, C. (2016). *Social Media Research: A Guide to Ethics*. [online] The University of Aberdeen. Available at: <https://www.utwente.nl/en/bms/research/forms-and-downloads/socialmediaresearchethics.pdf> [Accessed 25 Aug. 2025].

Tsormpatzoudi, P., Cir, I., Leuven, K., Preneel, B. and Cosic, E. (2018). *Collateral damage of Facebook Apps: an enhanced privacy scoring model*. [online] Available at: <https://eprint.iacr.org/2015/456.pdf> [Accessed 6 Sep. 2025].

Tunguz, B. (2019). *Big Five Personality Test*. [online] www.kaggle.com. Available at: <https://www.kaggle.com/datasets/tunguz/big-five-personality-test> [Accessed 20 Sep. 2025].

Tuovinen, S., Tang, X. and Salmela-Aro, K. (2020). Introversion and Social Engagement: Scale Validation, Their Interaction, and Positive Association with Self-Esteem. *Frontiers in Psychology*, [online] 11(590748), pp.1–11. doi:<https://doi.org/10.3389/fpsyg.2020.590748> [Accessed 3 Oct. 2025].

University of Glasgow (2025). *University of Glasgow - MyGlasgow - MyGlasgow Staff-Brand Toolkit - Resources and guides - UofG Social Media Guidelines - Social Media at UofG*. [online] Gla.ac.uk. Available at:

<https://www.gla.ac.uk/myglasgow/staff/brand toolkit/resources/socialmedia/social-media-uofg/> [Accessed 12 Oct. 2025].

Ur Rehman, I. (2019). *Facebook-Cambridge Analytica data harvesting: What you need to know*. [online] University of Nebraska - Lincoln. Available at: <https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=5833&context=libphilprac> [Accessed 22 Aug. 2025].

Urbansky, D. (2022). *World News API*. [online] Worldnewsapi.com. Available at: <https://worldnewsapi.com/> [Accessed 19 Oct. 2025].

von Bergen, H. and Diestel, R. (2025). Traits and tangles: An analysis of the Big Five paradigm by tangle-based clustering. *Journal of Mathematical Psychology*, [online] 125, p.102920. doi:<https://doi.org/10.1016/j.jmp.2025.102920> [Accessed 28 Sep. 2025].

W3Schools (2024). *Pandas Tutorial*. [online] www.w3schools.com. Available at: <https://www.w3schools.com/python/pandas/default.asp> [Accessed 21 Sep. 2025].

W3Schools (2025). *Matplotlib Tutorial*. [online] www.w3schools.com. Available at: https://www.w3schools.com/python/matplotlib_intro.asp [Accessed 22 Sep. 2025].

W3Schools (n.d.). *Python Machine Learning - K-means*. [online] www.w3schools.com. Available at: https://www.w3schools.com/python/python_ml_k-means.asp [Accessed 23 Sep. 2025].

Wagner, P. (2021). Data Privacy - The Ethical, Sociological, and Philosophical Effects of Cambridge Analytica. *SSRN Electronic Journal*. [online] doi:<https://doi.org/10.2139/ssrn.3782821> [Accessed 22 Aug. 2025].

Web.archive.org (2016). *SCL Group | Home*. [online] Archive.org. Available at: <https://web.archive.org/web/20161205184851/http://sclgroup.cc/?hg=0> [Accessed 2 Sep. 2025].

Wolski, M. and Gomolińska, A. (2020). Data meaning and knowledge discovery: Semantical aspects of information systems. *International Journal of Approximate Reasoning*, [online] 119, pp.40–57. doi:<https://doi.org/10.1016/j.ijar.2020.01.002> [Accessed 3 Aug. 2025].

Wong, J.C. (2019). *Hundreds of millions of Facebook records exposed on public servers – report*. [online] the Guardian. Available at: <https://www.theguardian.com/technology/2019/apr/03/facebook-data-public-servers-amazon> [Accessed 18 Sep. 2025].

Wong, J.C. and Lewis, P. (2018). Facebook gave data about 57bn friendships to academic. *The Guardian*. [online] 22 Mar. Available at: <https://www.theguardian.com/news/2018/mar/22/facebook-gave-data-about-57bn-friendships-to-academic-aleksandr-kogan> [Accessed 4 Sep. 2025].

Wong, J.C., Lewis, P. and Davies, H. (2018). *How academic at centre of Facebook scandal tried – and failed – to spin personal data into gold*. [online] the Guardian. Available at: <https://www.theguardian.com/news/2018/apr/24/aleksandr-kogan-cambridge-analytica-facebook-data-business-ventures> [Accessed 2 Sep. 2025].

Woods, L. (2024). *Disinformation and disorder: the limits of the Online Safety Act*. [online] Online Safety Act Network. Available at: <https://www.onlinesafetyact.net/analysis/disinformation-and-disorder-the-limits-of-the-online-safety-act/> [Accessed 9 Oct. 2025].

Zapier (2019). *The 9 Best Free Form Builders and Survey Tools*. [online] Zapier. Available at: <https://zapier.com/blog/best-free-survey-tool-form-app/> [Accessed 7 Sep. 2025].

Zinolabedini, D. and Arora, N. (2019). *RUNNING HEAD: FACEBOOK DATA SCANDAL 1 The Ethical Implications of the 2018 Facebook-Cambridge Analytica Data Scandal*. [online] Available at: <https://repositories.lib.utexas.edu/server/api/core/bitstreams/25af2643-9128-49d2-863e-cb5ff3aee9ce/content> [Accessed 27 Aug. 2025].

Zollo, F. and Quattrociocchi, W. (2018). *Misinformation Spreading on Facebook*. [online] Cham, Switzerland: Springer International Publishing, pp.177–193. Available at: https://doi.org/10.1007/978-3-319-77332-2_10 [Accessed 10 Aug. 2025].

Bibliography:

- Amnesty International (2024). *UK: X created a ‘staggering amplification of hate’ during the 2024 riots*. [online] Amnesty International UK. Available at: <https://www.amnesty.org.uk/press-releases/uk-x-created-staggering-amplification-hate-during-2024-riots> [Accessed 9 Oct. 2025].
- Bakir, V. (2020). Psychological Operations in Digital Political Campaigns: Assessing Cambridge Analytica’s Psychographic Profiling and Targeting. *Frontiers in Communication*, [online] 5. doi:<https://doi.org/10.3389/fcomm.2020.00067> [Accessed 18 Sep. 2025].
- Business & Human Rights Resource Centre (2025). *UK: Far-right riots allegedly fuelled by social media misinformation spread on X, Telegram, Instagram & Facebook - Business & Human Rights Resource Centre*. [online] Business & Human Rights Resource Centre. Available at: <https://www.business-humanrights.org/en/latest-news/uk-far-right-riots-allegedly-fuelled-by-social-media-misinformation-spread-on-x-telegram-instagram-facebook/> [Accessed 9 Oct. 2025].
- Channel 4 News Investigations Team (2018). *Revealed: Cambridge Analytica data on thousands of Facebook users still not deleted*. [online] Channel 4 News. Available at: <https://www.channel4.com/news/revealed-cambridge-analytica-data-on-thousands-of-facebook-users-still-not-deleted> [Accessed 5 Sep. 2025].
- Cheshire, T. and Doak, S. (2024). *Southport attack misinformation fuels far-right discourse on social media*. [online] Sky News. Available at: <https://news.sky.com/story/southport-attack-misinformation-fuels-far-right-discourse-on-social-media-13188274> [Accessed 9 Oct. 2025].
- Constantiou, I. and Kallinikos, J. (2015). New Games, New Rules: Big Data and the Changing Context of Strategy. *Journal of Information Technology*, [online] 30(1), pp.44–57. doi:<https://doi.org/10.1057/jit.2014.17> [Accessed 22 Aug. 2025].
- Denham, E. (2020). *RE: ICO investigation into use of personal information and political influence*. [online] Available at: https://ico.org.uk/media2/migrated/2618383/20201002_ico-o-ed-l-rtl-0181_to-julian-knight-mp.pdf [Accessed 22 Sep. 2025].
- Department for Science, Innovation & Technology (2025). *Online Safety Act: Explainer*. [online] GOV.UK. Available at: <https://www.gov.uk/government/publications/online-safety-act-explainer/online-safety-act-explainer> [Accessed 15 Oct. 2025].
- GeeksforGeeks (2020). *Pandas Introduction*. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/pandas/introduction-to-pandas-in-python/> [Accessed 21 Nov. 2025].
- Global News (2018). *Facebook CEO Mark Zuckerberg’s FULL testimony to U.S. congress members*. [online] YouTube. Available at: https://www.youtube.com/watch?v=YCQ_ZGxE2U4 [Accessed 14 Oct. 2025].

Haugen, F. (2021). *Statement of Frances Haugen*. [online] Commerce.senate.gov. Available at: <https://www.commerce.senate.gov/services/files/FC8A558E-824E-4914-BEDB-3A7B1190BD49> [Accessed 14 Sep. 2025].

Huang, K. and Krafft, P.M. (2024). Performing Platform Governance: Facebook and the Stage Management of Data Relations. *Science and engineering ethics*, [online] 30(2). doi:<https://doi.org/10.1007/s11948-024-00473-5> [Accessed 14 Sep. 2025].

ICO (2018). *Investigation into the use of data analytics in political campaigns*. [online] Information Commissioner's Office. Available at: <https://ico.org.uk/media2/migrated/2260271/investigation-into-the-use-of-data-analytics-in-political-campaigns-final-20181105.pdf> [Accessed 22 Sep. 2025].

Ingram, D. (2018). Factbox: Who is Cambridge Analytica and what did it do? *Reuters*. [online] 20 Mar. Available at: <https://www.reuters.com/article/technology/factbox-who-is-cambridge-analytica-and-what-did-it-do-idUSKBN1GW07F/> [Accessed 13 Oct. 2025].

Isaak, J. and Hanna, M.J. (2018). *User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection - IEEE Journals & Magazine*. [online] Ieee.org. Available at: <https://ieeexplore.ieee.org/document/8436400> [Accessed 1 Sep. 2025].

Jaiswal, A., Shah, A., Harjadi, C., Windgassen, E. and Washington, P. (2024). Ethics of the Use of Social Media as Training Data for Artificial Intelligence Models used for Digital Phenotyping: Commentary (Preprint). *JMIR Formative Research*, [online] 8, pp.e59794–e59794. doi:<https://doi.org/10.2196/59794> [Accessed 24 Aug. 2025].

Johns Hopkins University School of Medicine Biochemistry (n.d.). *What is Phenotyping? – The Department of Molecular & Comparative Pathobiology*. [online] Johns Hopkins University School of Medicine Biochemistry (JHUSOM). Available at: <https://mcp.bs.jhmi.edu/what-is-phenotyping/> [Accessed 24 Aug. 2025].

Klander, O. (2020). *Psychometric personality-based factors in (OCEAN) targeting*. [online] Oliver Klander. Available at: <https://www.klander.com/oliver-klander-blog/psychometric-personality-based-factors-in-ocean-targeting> [Accessed 3 Oct. 2025].

Kleinman, Z. (2018). Cambridge Analytica: The story so far. *BBC News*. [online] 21 Mar. Available at: <https://www.bbc.co.uk/news/technology-43465968> [Accessed 13 Oct. 2025].

LSE (n.d.). *Code of Research Conduct*. [online] Available at: <https://info.lse.ac.uk/staff/services/Policies-and-procedures/Assets/Documents/codResCon.pdf> [Accessed 12 Oct. 2025].

Ma, A. and Gilbert, B. (2019). *Facebook understood how dangerous the Trump-linked data firm Cambridge Analytica could be much earlier than it previously said. Here's everything that's happened up until now*. [online] Business Insider. Available at:

<https://www.businessinsider.com/cambridge-analytica-a-guide-to-the-trump-linked-data-firm-that-harvested-50-million-facebook-profiles-2018-3> [Accessed 13 Oct. 2025].

Meta (2025). *Supplemental Meta Platforms Technologies Privacy Policy | Meta*. [online] Meta.com. Available at: <https://www.meta.com/gb/legal/privacy-policy/> [Accessed 4 Aug. 2025].

Moreno, M.A., Kota, R., Schoohs, S. and Whitehill, J.M. (2013). The Facebook Influence Model: A Concept Mapping Approach. *Cyberpsychology, Behavior, and Social Networking*, [online] 16(7), pp.504–511. doi:<https://doi.org/10.1089/cyber.2013.0025> [Accessed 24 Aug. 2025].

Müller, O. and Degaetano-Ortlieb, S. (2025). *Embedded Personalities: Word Embeddings and the ‘Big Five’ Personality Model*. [online] pp.205–215. Available at: <https://aclanthology.org/2025.latechclfl-1.18.pdf> [Accessed 6 Oct. 2025].

Oudin, A., Maatoug, R., Bourla, A., Ferreri, F., Bonnot, O., Millet, B., Schoeller, F., Mouchabac, S. and Adrien, V. (2023). Digital Phenotyping: Data-Driven Psychiatry to Redefine Mental Health. *Journal of Medical Internet Research*, [online] 25(1), p.e44502. doi:<https://doi.org/10.2196/44502> [Accessed 24 Aug. 2025].

Perlroth, N., Frenkel, S. and Shane, S. (2018). Facebook Executive Planning to Leave Company Amid Disinformation Backlash. *The New York Times*. [online] 19 Mar. Available at: <https://www.nytimes.com/2018/03/19/technology/facebook-alex-stamos.html> [Accessed 13 Oct. 2025].

Research, S. (2024). *Sage Research Methods Community*. [online] Sage Research Methods Community. Available at: <https://researchmethodscommunity.sagepub.com/blog/informed-consent-online-research> [Accessed 24 Aug. 2025].

Romano, A. (2018). *The Facebook data breach wasn’t a hack. It was a wake-up call*. [online] Vox. Available at: <https://www.vox.com/2018/3/20/17138756/facebook-data-breach-cambridge-analytica-explained> [Accessed 13 Oct. 2025].

Schechner, S. (2019). *Eleven Popular Apps That Shared Data With Facebook*. [online] Available at: <https://www.wsj.com/articles/eleven-popular-apps-that-shared-data-with-facebook-11551055132> [Accessed 6 Sep. 2025].

Solnik, C. (2018). *Cambridge Analytica filing for bankruptcy*. [online] Long Island Business News. Available at: <https://libn.com/2018/05/02/cambridge-analytica-filing-for-bankruptcy/> [Accessed 13 Sep. 2025].

The European Parliament (2020). *Texts adopted - The use of Facebook users’ data by Cambridge Analytica and the impact on data protection - Thursday, 25 October 2018*. [online] www.europarl.europa.eu. Available at:

https://www.europarl.europa.eu/doceo/document/TA-8-2018-0433_EN.html [Accessed 14 Oct. 2025].

The New York Times (2018). *How Cambridge Analytica Exploited the Facebook Data of Millions* | NYT. YouTube. Available at: <https://www.youtube.com/watch?v=mrnXv-g4yKU> [Accessed 17 Sep. 2025].

Wade, M. (2022). *Psychographics: the behavioural analysis that helped Cambridge Analytica know voters' minds*. [online] www.imd.org. Available at: <https://www.imd.org/research-knowledge/technology-management/articles/psychographics-the-behavioural-analysis-that-helped-cambridge-analytica-know-voters-minds/> [Accessed 22 Sep. 2025].

Wang, N. (2012). Third-party applications' data practices on facebook. [online] pp.1399–1404. doi:<https://doi.org/10.1145/2212776.2212462> [Accessed 13 Oct. 2025].

Wong, J.C. (2018). 'It might work too well': the dark art of political advertising online. *The Guardian*. [online] 19 Mar. Available at: <https://www.theguardian.com/technology/2018/mar/19/facebook-political-ads-social-media-history-online-democracy> [Accessed 18 Sep. 2025].

Appendix:

1. Kaggle Dataset for OCEAN traits for UK participants

The dataset utilised in this study comprises data exclusively collected from participants based in the United Kingdom. It serves to assess psychological profiles using the OCEAN personality traits model. The complete dataset, available for review on GitHub, is provided in both Microsoft Excel (.xlsx) and Comma-Separated Values (.csv) formats. It is important to note that the dataset comprises 66,487 rows and 3,324,350 data points. Due to its substantial size, users may have difficulty loading or displaying the entire dataset unless they access the raw file or download it (Tunguz, 2019).

The GitHub file names for the data file are **data-final-CLEAN-GB-Data-GitHub.xlsx** and **data-final-CLEAN-GB-Data-GitHub.csv**

GitHub: <https://github.com/abmiah/msc-dissertation-appendix-files>

2. K-Means Python Code

The full Colab code and .csv file with only the averages are available in the GitHub repository, named **KMean.ipynb** and **data-final-CLEAN_GB_Data_for_KMean.csv**. Below, you'll find the code stored both on GitHub and the original code created in Google Colab.

Import the CSV filename. Please note that if the terminal times out, you will need to re-upload the CSV data file.

GitHub: <https://github.com/abmiah/msc-dissertation-appendix-files/>

Colab: https://colab.research.google.com/drive/1TLoXbjFcV3tgT6veZe_ps44JOPXLIVE1

Below is the code sample taken directly from Colab IDE.

The script starts by importing the necessary libraries to enable k-means clustering graph.

These libraries are essential for data analysis, visualisation, and clustering. Pandas provides data import, cleaning, and organisation for large datasets through DataFrames. **matplotlib.pyplot** generates clear visuals, such as histograms and scatter plots, to help interpret data. **sklearn.cluster** import **KMeans** performs K-Means clustering, grouping data points by features to identify patterns in the dataset.

```
# Import the necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
```

This script loads a cleaned CSV into a pandas DataFrame from 'data-final-CLEAN_GB_Data_for_KMean.csv' for analysis, including statistical tasks, visualisation, and clustering Python.

```
# Import the CSV filename. Please note that if the terminal times out, you will need to re-upload the CSV data file.
data = pd.read_csv('/content/data-final-CLEAN_GB_Data_for_KMean.csv')
```

This script extracts specific columns corresponding to the Big Five personality traits, **Openness**, **Conscientiousness**, **Extraversion**, **Agreeableness**, and **Neuroticism**, from the dataset. It assigns these column names to the list trait_columns and selects them from data to create a focused subset (X) for subsequent statistical or machine learning analysis involving personality traits evaluation.

```
# Headline should match the headline within the csv file
trait_columns = ['Openness', 'Conscientiousness', 'Extraversion', 'Agreeableness',
'Neuroticism']
X = data[trait_columns]
```

This script uses the k-means clustering algorithm to categorise data points in X into 7 clusters. It sets a fixed random seed to ensure reproducibility and runs the

algorithm multiple times for stability before assigning each data row to its respective cluster label.

```
# Run k-means (edit the "k=xxx" to get different values)
k = 7
kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
data['Cluster'] = kmeans.fit_predict(X)
```

This script displays the first 10 dataset rows for quick review of structure, format, and content after loading. It helps verify columns, data types, and spot issues like missing or incorrectly data value.

```
# Print the first 10 data (noted that the index starts at "0")
print(data.head(10))
```

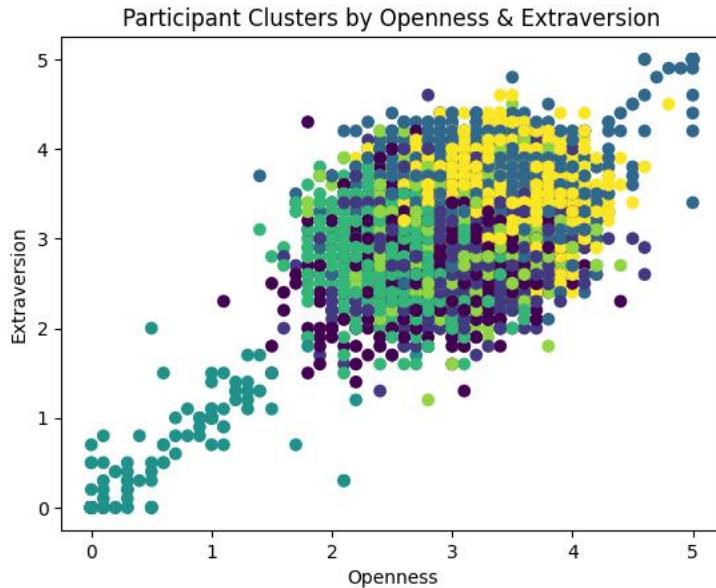
| | Openness | Conscientiousness | Extraversion | Agreeableness | Neuroticism | \ |
|---|----------|-------------------|--------------|---------------|-------------|---|
| 0 | 3.0 | 1.0 | 3.1 | 3.2 | 3.3 | |
| 1 | 2.9 | 2.6 | 2.8 | 2.8 | 3.1 | |
| 2 | 2.6 | 2.7 | 3.2 | 2.7 | 3.1 | |
| 3 | 2.8 | 2.8 | 2.9 | 2.5 | 3.3 | |
| 4 | 3.0 | 3.6 | 2.2 | 3.3 | 3.2 | |
| 5 | 2.7 | 3.4 | 3.3 | 3.1 | 3.0 | |
| 6 | 3.1 | 2.6 | 3.1 | 2.6 | 2.7 | |
| 7 | 3.5 | 2.6 | 3.4 | 2.6 | 3.0 | |
| 8 | 2.7 | 3.5 | 3.0 | 3.4 | 3.1 | |
| 9 | 3.0 | 4.1 | 2.9 | 3.0 | 3.4 | |

| | Cluster |
|---|---------|
| 0 | 1 |
| 1 | 0 |
| 2 | 0 |
| 3 | 0 |
| 4 | 4 |
| 5 | 5 |
| 6 | 0 |
| 7 | 0 |
| 8 | 5 |
| 9 | 5 |

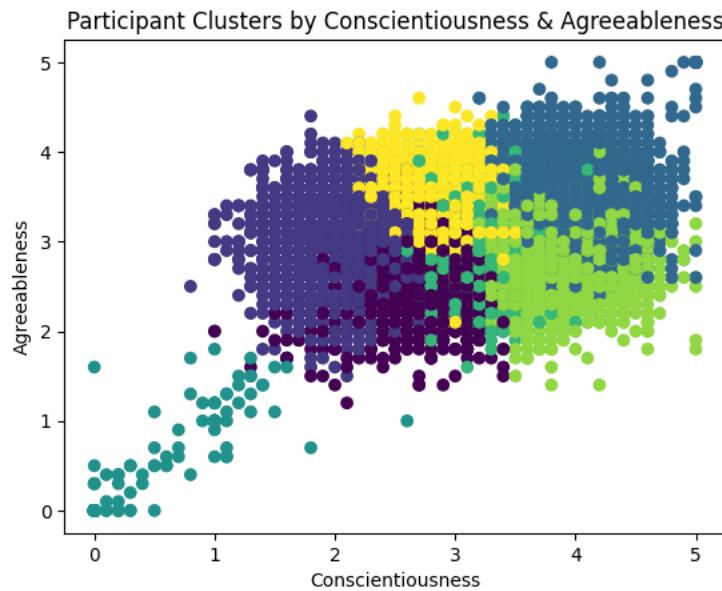
This code visualises participant clustering based on **Openness** and **Extraversion** by plotting a scatter plot where each point represents an individual. The point's colour indicates the cluster, showing subgroup distribution and potential patterns between the traits sample. This

code is repeated with **Conscientiousness** and **Agreeableness** and then with **Agreeableness** and **Neuroticism**.

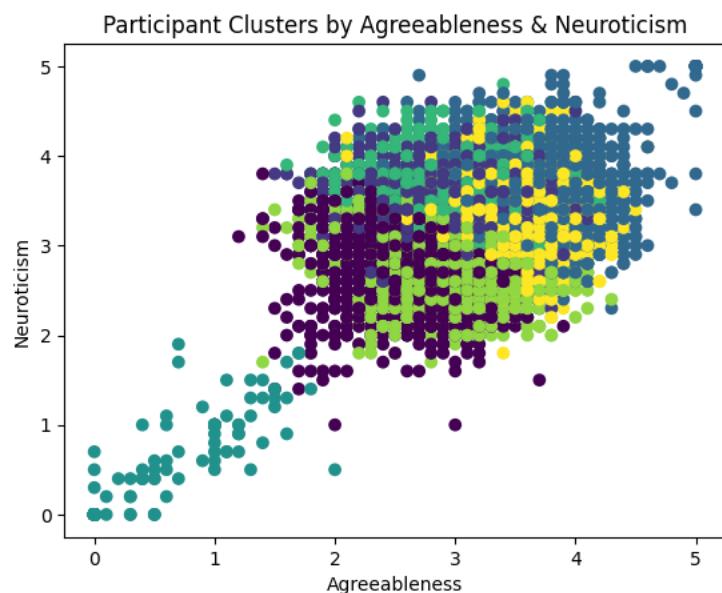
```
plt.scatter(data['Openness'], data['Extraversion'], c=data['Cluster'], cmap='viridis')
plt.xlabel('Openness')
plt.ylabel('Extraversion')
plt.title('Participant Clusters by Openness & Extraversion')
plt.show()
```



```
plt.scatter(data['Conscientiousness'], data['Agreeableness'], c=data['Cluster'], cmap='viridis')
plt.xlabel('Conscientiousness')
plt.ylabel('Agreeableness')
plt.title('Participant Clusters by Conscientiousness & Agreeableness')
plt.show()
```



```
plt.scatter(data['Agreeableness'], data['Neuroticism'], c=data['Cluster'], cmap='viridis')
plt.xlabel('Agreeableness')
plt.ylabel('Neuroticism')
plt.title('Participant Clusters by Agreeableness & Neuroticism')
plt.show()
```



This script shows cluster centres (centroids) from **K-Means**, representing average trait scores for each group to interpret cluster differences across personality traits, presented as a readable DataFrame for easy comparison analysis.

```
# Print cluster centres
print("Cluster centres (average trait scores per group): ","\n")
# Print the table and formatted to make it readable
```

```
print(pd.DataFrame(kmeans.cluster_centers_, columns=trait_columns))
```

Cluster centres (average trait scores per group):

| | Openness | Conscientiousness | Extraversion | Agreeableness | Neuroticism |
|---|----------|-------------------|--------------|---------------|-------------|
| 0 | 2.994491 | 2.742674 | 2.982321 | 2.772797 | 2.927750 |
| 1 | 3.008246 | 2.210734 | 3.050879 | 3.014466 | 3.413152 |
| 2 | 3.063043 | 3.887495 | 3.336722 | 3.418302 | 3.531913 |
| 3 | 0.361607 | 0.322768 | 0.345982 | 0.336161 | 0.311161 |
| 4 | 2.781251 | 3.220750 | 2.967832 | 3.062127 | 3.517167 |
| 5 | 2.997422 | 3.689652 | 3.117689 | 2.956861 | 2.989889 |
| 6 | 3.215027 | 2.954407 | 3.333698 | 3.271563 | 3.369383 |

3. Table 4 of OCEAN Traits with their score values and psychological profiles description.

| Trait | High Score Description | Low Score Description |
|-------------------|---|--|
| Openness | Tolerant, imaginative, emotionally sensitive, independent. | Superficial, unrefined, and lacking insight routine based. |
| | This trait characterises an individual as possessing open-mindedness, empathy, and a strong sense of values directed towards others and their environment, complemented by emotional awareness. Such individuals typically demonstrate a broad appreciation for diverse emotional perspectives and exhibit openness to creative ideas (Chacón et al., 2024; Ng et al., 2021). | They tend to lack a depth of thought and sensitivity and awareness, which indicates a low level of open-mindedness in such areas as imagination, insight and general intellectual curiosity, which will indicate a person with simplistic thinking, a lack of awareness and only focused on the appearance on a surface level rather than a more profound understanding (Diener et al., 2019). |
| Conscientiousness | Self-disciplined, motivated, organised, and considerate beforehand acting. | Often careless, untrustworthy, sloppy behaviour and often disorganised. |
| | A person exhibiting a high level of conscientiousness typically shows enhanced self-regulation and a strong focus on achieving goals. Such individuals usually | This trait is primarily characterised by individuals displaying irresponsibility and low levels of conscientiousness. Their behavioural patterns are |

| | | |
|---------------|--|---|
| | engage in thorough planning and display deliberate foresight before making decisions, often thoughtfully considering their actions (Jackson et al., 2010). | often associated with carelessness, a lack of attentiveness to detail during task supervision, and dishonest or deceptive behaviours, which are influenced by their cognitive and social factors (Chauliac et al., 2023). |
| Extraversion | Outgoing, warm, enthusiastic about new experiences, and assertive. | Introverted, passive, and reserved. |
| | These individuals exhibit a high level of energy and possess a broad social network, which enables them to feel comfortable in group settings. They are characterised by sociability, a propensity for seeking new challenges, and a penchant for adventure and enthusiasm. Furthermore, they tend to be assertive, often to stimulate new experiences or interactions (Opoku et al., 2023). | Individuals exhibiting traits of introversion, passivity, and reserved behaviour typically demonstrate low levels of extraversion. Consequently, they tend to be more solitary or withdrawn within social contexts, often appearing quiet or timid and displaying reduced assertiveness during interactions. Such individuals generally prefer self-reflection over external stimulation. This personality trait is less commonly observed in leadership positions or highly social roles and is more characteristic of personal introspection (Tuovinen et al., 2020). |
| Agreeableness | Modest, trustful, cooperative, and straightforward. | Selfish, hostile, and inattentive to others. |
| | Individuals exhibiting this trait demonstrate heightened levels of compassion and empathy, actively engaging in cooperative interactions with others who manifest qualities such as trust, humility, and frankness. These individuals are frequently perceived as benevolent and reliable, often demonstrating | Individuals exhibiting this trait demonstrate egocentrism and a persistent detachment from the broader social environment. Such individuals tend to be emotionally distant, distrustful, and display a notable lack of empathetic concern towards others. These characteristics may manifest in behaviours |

| | | |
|-------------|--|--|
| | <p>concern for the welfare of others and striving to establish harmonious social relationships. Their interpersonal interactions are typically characterised by low aggression and high levels of understanding (Lim, 2025).</p> | <p>characterised by selfishness and hostility. Empirical evidence suggests that these individuals often score lower on measures of agreeableness and exhibit significant correlations with the 'dark triad' personality traits, including narcissism, Machiavellianism, and psychopathy. The cold and self-centred behaviours associated with this trait can result in contentious and callous interactions within social contexts, thereby adversely impacting interpersonal relationships and broader social outlooks (Anderson et al., 2020).</p> |
| Neuroticism | Hostile, angry, impulsive, anxious, and self-conscious. | Calm and self-possessed even-tempered. |
| | <p>Individuals exhibiting high levels of Neuroticism tend to manifest persistent negative emotional traits, including heightened anxiety, self-consciousness, anger, impulsivity, and occasional hostility towards others. They are often characterised by impaired impulse control and increased reactivity to stress (Dam et al., 2021).</p> | <p>Individuals characterised by low levels of Neuroticism are commonly described as possessing emotional stability. Such individuals typically exhibit calmness, emotional equilibrium, composure, and an enhanced ability to manage stress. Additionally, they generally exhibit fewer mood fluctuations, reduced anxiety, and diminished emotional reactivity (Lim, 2025).</p> |

Table 4: Provides a succinct overview of personality traits based on their respective high and low score values, which collectively outline an individual's psychological profile (Lim, 2025; Nasello et al., 2023; Sutton, 2025).

4. Mitigation Hypothesis Code.

```

◆ main.py > ⌂ run_pipeline
1  """
2 Main pipeline script to load content items, enrich with OCR, fetch news articles,
3 and match content items with news articles. This is also the entry point of the application.
4 """
5
6 """
7 The initial import statements bring in necessary modules and functions from other parts of the application.
8 These imports are essential for the pipeline to function correctly, as they provide access to database operations,
9 OCR processing, news fetching, and matching logic. Some of the imported modules also include data models and configuration
10 settings, and others are imported for future use or to maintain consistency across the application.
11 The from app import db, ocr, news, match statement imports the db, ocr, news, and match modules from the app package,
12 which is based within the same project structure.
13 """
14 from app import db, ocr, news, match
15 from app.models import ContentItem
16 from app.config import NEWS_API_KEY
17
18 """
19 The run_pipeline() function orchestrates the entire process of loading content items from the database,
20 enriching them with OCR data, fetching related news articles, and matching the content items with the news articles.
21 This function also handles the display of results, including matched items and unmatched items that may indicate
22 potential misinformation. """
23
24 """ When the function is executed, it performs the following steps:
25 1. Loads content items from the database using db.fetch_content_items().
26 2. Enriches the loaded items with OCR data using ocr.enrich_with_ocr().
27 3. Fetches related news articles and matches them with the content items using match.match_items_with_news().
28 4. Displays the matched items along with their similarity scores and article details."""
29
30 def run_pipeline():
31     print("Loading items from DB...")
32     items = db.fetch_content_items()
33     print(f"Loaded {len(items)} items.")
34
35     print("Enriching with OCR (if files exist...)")
36     items = ocr.enrich_with_ocr(items)
37
38     print("Fetching related news and matching from all enabled APIs...")
39     matches = match.match_items_with_news(items, news.fetch_news_articles)
40
41     # Track which items have matches (by item object, not filename)
42     matched_items = {id(m.item) for m in matches}
43
44     print("*"*80)
45     print(f"\nFound {len(matches)} matches:")
46     print("*"*80)
47     for i, m in enumerate(matches, 1):
48         print(f"\nMatch #{i} | score={m.similarity:.2%}")
49         print(f" - File: {m.item.file_name}")
50         print(f" - Content: {m.item.content_info[:80]}")
51         if m.item.ocr_text:
52             print(f" - OCR Text: {m.item.ocr_text[:80]}")
53         print(f" - Article: {m.article.title}")
54         print(f" - Source: {m.article.source}")
55         print(f" - URL: {m.article.url}")
56
57     # Report items without matches as potential false information
58     """ For any content items that did not find a match in the news articles,
59     the function reports them as potential false information. This is done by checking
60     which items were not included in the matched_items set and printing their details
61     along with a cautionary message. """
62     unmatched_items = [item for item in items if id(item) not in matched_items]
63     if unmatched_items:
64         print("\n*"*80)
65         print(f"UNMATCHED ITEMS ({len(unmatched_items)})")
66         print("*"*80)
67         for item in unmatched_items:
68             print(f"\n[A] File: {item.file_name}")
69             print(f" Content: {item.content_info[:80]}")
70             print(f" Status: The information from this post cannot be verified [B], please be cautious of potential misinformation.")
71             print(f"\n-*80")
72
73
74 """
75 The if __name__ == '__main__': block ensures that the main() function is called only when the script is run directly,
76 and not when it is imported as a module in another script. This allows users to see how the script works and what results
77 it produces. This is also common practice in Python programming. """
78 if __name__ == '__main__':
79     run_pipeline()

```

Figure 14: A sample of the Python code. The Main.py is the main script that runs the code in the terminal. Within the README.md file below, snippets of the code will also be included.

GitHub: <https://github.com/abmiah/msc-dissertation-code>

Video Demo: <https://www.youtube.com/watch?v=aCb6T-xMF7w>

The README.md file provides an overview of the mitigation code.

1. Introduction

This README outlines a hypothetical mitigation strategy to counter misinformation on social media, with a particular focus on Facebook. Over the preceding year, there have been numerous instances of misinformation that have influenced user behaviour and targeted specific psychological predispositions. Of particular interest is the susceptibility of individuals exhibiting high levels of Neuroticism to microtargeting campaigns. Such individuals tend to display behavioural traits characterised by persistent negative emotions, including anxiety, self-consciousness, anger, impulsivity, and hostility. Additionally, they often demonstrate poor impulse control and heightened stress reactivity. Conversely, individuals with low levels of Neuroticism generally exhibit traits of calmness, self-possession, and even-temperedness (Dam et al., 2021; Lim, 2025; Sutton, 2025) During the preliminary phase of the written study involving UK participants, the results revealed that individuals with elevated levels of Neuroticism exhibit increased susceptibility to microtargeted advertisements, primarily due to heightened emotional fear responses that subsequently trigger behavioural reactions. These findings align with existing research supported by Bakir (2020) and Prichard (2021), which emphasises the vulnerability of specific populations to digital influence. The dissemination of misinformation presents significant implications in the digital era, especially given the rapid and pervasive nature of information spread facilitated by social media platforms such as Facebook, which boasts over three billion active users (Dixon, 2025c). Facebook also collects extensive data on its users, including behavioural patterns, social network connections, product usage, transactional activities on the platform, and analyses of user interactions within their networks, primarily for targeted advertising (Meta, 2022). Companies and governmental agencies can utilise Facebook's advertising platform, leveraging micro-targeting to customise ads based on specific demographic parameters, such as location and gender. Nonetheless, the ethical implications of such targeted advertising remain a subject of debate, particularly given the lack of oversight concerning the potential dissemination of misinformation. Furthermore, it has been alleged that entities like Cambridge Analytica exerted influence over various political campaigns, including the deployment of targeted advertisements that purportedly contributed to Donald J. Trump's electoral victory in the 2016 United States presidential election, wherein he secured 304 electoral votes compared to Hillary Clinton's 227. (History.com Editors, 2018; The New York Times, 2017) Alongside other political campaigns from Australia, several African nations, as well as Mexico, Brazil, India, and Malaysia, have employed digital microtargeting strategies for purposes of political and social manipulation (BBC News, 2018b; Netflix, 2019; Nyabola, 2019; SBS News, 2018). This indicates that microtargeting and data-driven misinformation are not restricted to specific regions or cultural contexts but are progressively emerging as global trends.

Please Note: Code execution is purely hypothetical

This coding framework is purely hypothetical; consequently, it is not possible to conduct a live scan on Facebook's actual system due to the lack of academic access. The code

references previous advertisements on Facebook that have raised ethical concerns related to misinformation.

2. Current Situation as of the present analysis.

The UK's Online Safety Act 2023 seeks to establish a regulatory framework aimed at controlling and mitigating harmful online content, with a particular emphasis on addressing the proliferation of misinformation across social media platforms. (Legislation.Gov.UK, 2023), misinformation contributed to a violent riot that took place on July 29, 2024, sparked by claims that a Muslim asylum seeker was involved in a brutal attack at a dance studio in Southport. The spread of false information on social media reached millions of users. The individual was later identified as a British national with no links to Islam nor any status as an illegal immigrant. While this does not implicate the individual in the violent act, the ensuing demonstrations resulted in clashes with law enforcement personnel, with some demonstrators blaming asylum seekers and the Muslim community. It can be argued that social media platforms failed to adequately address the dissemination of misinformation; in fact, there are claims that Facebook's algorithms may have even amplified such content due to high engagement levels (BBC Bitesize, 2024; Benesch, 2021; Fung, 2024; Kiderlin, 2024; Mohamed, 2024; Shah, 2024). According to Woods (2024), the legislation itself fails to directly address the issue of misinformation, primarily due to the absence of clearly defined criteria for criminality and intent.

3. Methodology

A software tool could be integrated into social media platforms such as Facebook and other relevant networks to enhance the detection of misinformation. This approach would leverage UK-specific initiatives to combat disinformation by cross-referencing advertisements and viral posts via a news API. Nonetheless, this endeavour faces several challenges, including issues of platform confidentiality and the necessity for collaboration between platform providers and government agencies. The primary objective is to mitigate the spread of misinformation by alerting users to verify content, providing credible sources, and assisting in classifying information as factual or otherwise.

4. Mitigation Hypothesis

To effectively combat the proliferation of misinformation, it is imperative to develop an automated, independent third-party tool. Such a tool should be seamlessly integrated into social media platforms to facilitate the verification of advertisements' authenticity and enable the identification of potential misinformation prior to its dissemination. This approach aims to enhance the integrity of information circulated within digital environments, supporting stakeholders such as users, researchers, and regulatory authorities. This research presents a modular, Python-based framework designed to address the pervasive issue of misinformation within online advertising. The system employs advanced Optical Character Recognition (OCR) (Lee, 2022) techniques to analyse visual media, such as images and videos, to extract textual content. This extracted data is systematically stored within a structured database, facilitating efficient retrieval and analysis. Subsequently, the framework utilises public news APIs to cross-reference the extracted information against reputable news sources. By

quantifying the similarity between ad content and verified journalistic articles, the system can identify and flag potentially unverified or false information. In a simulated environment, the methodology involves analysing local media files from a user's Facebook domain to extract and compare textual content with established news narratives. The modular architecture of the system allows for seamless integration of additional APIs, scalability to accommodate diverse data sources, and adaptability to real-world datasets. This approach contributes to broader efforts to combat misinformation in digital advertising and serves as a foundational tool for research on digital literacy, cybersecurity, and policy development.

5. Facebook Ad Content Verification System

This Python-based system is designed to detect and authenticate Facebook advertisements through comparative analysis of their content against reputable news sources. By cross-referencing advertisement material with verified news articles obtained via various news APIs, the system aims to identify potential misinformation.

6. Overview of the code structure

1. This process involves applying Optical Character Recognition (OCR) technology to analyse images and videos sourced from Facebook advertisements.
2. It systematically extracts textual information from these media files and performs.
3. Cross-referencing with multiple news API sources.
4. This methodology aims to identify potential misinformation by verifying the extracted content against credible news articles.

7. Code output

This program, upon execution, produces a structured output within the IDE terminal across three distinct stages. Initially, it retrieves content items from an SQLite database and incorporates Optical Character Recognition (OCR)-extracted text derived from image files. Subsequently, it performs queries across multiple news APIs, including NewsAPI.org (2019), NewsData (n.d.), WorldNewsAPI (Urbansky, 2022) and TheNewsAPI (2025), to identify articles corresponding to the extracted keywords. During this process, it also displays the queried sources alongside the number of articles retrieved from each. Finally, the program outputs the results divided into two categories: **matched content**, which includes similarity scores, titles, sources, and verification URLs, and **unmatched content**, which highlights items that could not be verified, thereby indicating potential misinformation. Each matched item is accompanied by a confidence percentage (e.g., 20%), whereas unmatched items feature warning indicators. The overall output is designed to facilitate ease of review for manual inspection, detailed analysis, or integration into automated workflows. Additionally, diagnostic messages, such as API errors or file warnings, are provided to assist with troubleshooting during execution.

8. Characteristics of the Code

The codebase exhibits a modular architectural design that delineates concerns across various components, including database management, OCR processing, news aggregation, and similarity assessment. Utilising Python's dataclasses, the system enforces type safety within data models, thereby fostering clear and maintainable interfaces between modules. The OCR component employs the Tesseract OCR engine, augmented with preprocessing techniques such as Contrast Limited Adaptive Histogram Equalization (CLAHE) (more information on CLAHE below) and adaptive thresholding, to improve recognition accuracy under conditions of faint or low-contrast text. The news aggregation mechanism integrates multiple sources concurrently, leveraging the parallel querying of diverse news APIs to enhance resilience against rate limiting and service outages. Similarity matching is implemented via Python's difflib.SequenceMatcher, facilitating efficient character-based comparisons and subsequent extraction of the most salient keywords, after filtering out common stop words. Metadata associated with media assets and extracted textual content are persistently stored within an SQLite database, enabling comprehensive analysis. The system architecture supports both autonomous scripts (e.g., database.py, facebookAd.py, newsAPI.py) and a consolidated pipeline (main.py), thereby providing operational flexibility. Configuration parameters, including thresholds and API credentials, are centralised within the app/config.py module, simplifying adjustments and tuning. Overall, the design emphasises extensibility, allowing seamless integration of new APIs, advanced matching algorithms, or enhanced logging frameworks with minimal modifications to existing code. This architectural approach underscores the system's adaptability and robustness for scalable, maintainable information processing.

Features of the code

- **Multi-source News Verification:** Supports multiple news APIs (NewsAPI.org, NewsData.io, WorldNewsAPI)
- **OCR Text Extraction:** Uses Tesseract OCR with preprocessing for enhanced text detection
- **Intelligent Matching:** Similarity-based matching between ad content and news articles
- **Database Storage:** SQLite database for storing and managing content
- **Modular Architecture:** Clean separation of concerns with dedicated modules
- **Ad Keyword Detection:** Identifies ads by detecting keywords like “Sponsored” or “Advertisement”

What is CLAHE? A quick overview

Contrast Limited Adaptive Histogram Equalisation (CLAHE) represents an advanced iteration of Adaptive Histogram Equalisation (AHE), specifically developed to mitigate the risk of over-enhancement in images exhibiting a certain degree of noise (Joseph et al., 2017).

Image noise refers to random variations in brightness and colour that manifest as speckles and textures, thereby compromising the clarity and detail of the image, analogous to film grain (Schulz, 2013). The CLAHE algorithm operates by partitioning the image into non-overlapping contextual regions or tiles. Instead of globally adjusting the entire image, the algorithm enhances each individual tile independently before seamlessly blending these enhanced regions to produce a uniformly processed image (Joseph et al., 2017).

9. System Requirements

Note: Commands differ slightly across macOS, Windows, and Linux—follow the OS-specific steps below.

- **Operating System:** macOS, Linux, or Windows (macOS used for development)
- **Python Version:** 3.8 or higher
- **Tesseract OCR engine:** Install at the OS level (not via pip)
 - macOS: brew install tesseract
 - Ubuntu/Debian: sudo apt-get install tesseract-ocr
 - Windows: Download from [Tesseract releases](#) and add to PATH
- **Python Packages:** pytesseract, opencv-python, Pillow, numpy, requests, newsapi-python
 - Install: pip install -r requirements.txt
- **Internet Connection:** Required for news API queries
- **API Keys:** Register and configure at least one provider (NewsAPI.org, NewsData.io, WorldNewsAPI; TheNewsAPI optional)
- **Media Folder:** Ensure image_and_video_directory/ contains files to scan
- **Disk Space:** Sufficient for media files and the SQLite database
- **RAM:** 2GB+ typically sufficient for small images; larger media may require more. Execute all commands from the project root. Verify Tesseract is on your PATH, and confirm installation with tesseract --version before running the pipeline.

10. Clone repository

You can clone the initial code from the GitHub repository using the following link:
<https://github.com/abmiah/msc-dissertation-code>

11. Running the script

- This codebase was developed in Visual Studio Code (VS Code). If you run commands in VS Code, output appears in the integrated terminal.
- Clone or navigate to the project directory.
- Install Tesseract OCR for your OS (see System Requirements above).
- Create and activate a Python virtual environment:
 - macOS/Linux:

- python3 -m venv .venv
 - source .venv/bin/activate
- Windows:
 - python -m venv .venv
 - .venv\Scripts\activate
- Install Python dependencies:
 - pip install -r requirements.txt
- Register at least one news API key (see Prerequisites) and add it to newsAPI.py.
- Ensure your media files are in the image_and_video_directory/ folder.
- Initialize the database and scan files:
 - python database.py
- Run the main verification pipeline:
 - python main.py
- Optional diagnostics or OCR testing:
 - python newsAPI.py (standalone news matching)
 - python facebookAd.py (OCR/ad keyword detection) All output, including results and error messages, will be shown in your terminal. When using VS Code, the integrated terminal provides the best experience.

12. Python file and Methods Overview

main.py

The script uses app.news.fetch_news_articles to retrieve potentially relevant articles and employs app.match.match_items_with_news to identify content-to-article matches. It then presents clear results, such as similarity scores, sources, and URLs. Items lacking matches are marked as potentially unverified. This process integrates configuration settings from app.config, data models from app.models, and several supporting modules. Use this approach for thorough end-to-end verification of the database content.

```
"""
Main script to load content items, enrich with OCR, fetch news articles,
and match them with news articles. This serves as the application entry point.

"""

"""

The initial import statements are crucial as they bring in the necessary modules and functions from other parts of
the
application. These imports ensure that the pipeline operates correctly by providing access to various
functionalities
such as database operations, OCR processing, news fetching, and matching logic. Some of the imported
modules also include
data models and configuration settings, while others are imported for future use or to maintain consistency
throughout
```

the application. Specifically, the statement `from app import db, ocr, news, match` imports the `db`, `ocr`, `news`, and

`match` modules from the app package, which is organized within the same project structure.

"""

```
from app import db, ocr, news, match
from app.models import ContentItem
from app.config import NEWS_API_KEY
```

"""

The `run_pipeline()` function coordinates the complete process of loading content items from the database, enhancing them with OCR data, retrieving related news articles, and matching these content items with the news articles. Additionally, this function manages the display of results, which includes both matched items and unmatched items that could suggest potential misinformation.

"""

"""

When the function is executed, it performs the following steps:

1. Loads content items from the database using db.fetch_content_items().
2. Enriches the loaded items with OCR data using ocr.enrich_with_ocr().
3. Fetches related news articles and matches them with the content items using match.match_items_with_news().
4. Displays the matched items along with their similarity scores and article details.
5. Identifies and reports any unmatched items as potential false information.

"""

```
def run_pipeline():
```

""" Orchestrates the entire verification pipeline: load data, enrich with OCR, fetch news, and match. """

```
print("Loading items from DB...")
```

```
items = db.fetch_content_items()
```

```
print(f"Loaded {len(items)} items.")
```

```
print("Enriching with OCR (if files exist)...")
```

```
items = ocr.enrich_with_ocr(items)
```

```
print("Fetching related news and matching from all enabled APIs...")
```

```
matches = match.match_items_with_news(items, news.fetch_news_articles)
```

```
# Track which items have matches (by item object, not filename)
```

```
matched_items = {id(m.item) for m in matches}
```

```

print("*80)
print(f"\nFound {len(matches)} matches:")
print("*80)
for i, m in enumerate(matches, 1):
    print(f"\nMatch #{i} | score={m.similarity:.2%}")
    print(f" - File: {m.item.file_name}")
    print(f" - Content: {m.item.content_info[:80]}")
    if m.item.ocr_text:
        print(f" - OCR Text: {m.item.ocr_text[:80]}")
    print(f" - Article: {m.article.title}")
    print(f" - Source: {m.article.source}")
    print(f" - URL: {m.article.url}")

# Report items without matches as potential false information
"""

For any content items that do not match any news articles, the function identifies them as potential
false information. This is accomplished by checking which items are absent from the matched_items set
and printing their details along with a cautionary message.

"""
unmatched_items = [item for item in items if id(item) not in matched_items]
if unmatched_items:
    print("\n{*80}")
    print(f"UNMATCHED ITEMS ({len(unmatched_items)}):")
    print("*80)
    for item in unmatched_items:
        print(f"\n⚠️ File: {item.file_name}")
        print(f" Content: {item.content_info[:80]}")
        print(f" Status: The information from this post cannot be verified[], please be cautious of potential
misinformation.")
    print(f"\n{-*80}")

"""

The `if __name__ == '__main__':` block ensures the `main()` function runs only when the script is executed
directly,
not when imported as a module. This practice is common in Python programming, allowing users to see how the
script works.
"""

if __name__ == '__main__':

```

```
run_pipeline()
```

database.py

This utility script sets up and populates the fbContentType.db Sqlite database with placeholder data and optionally scans media from image_and_video_directory/. It handles creating the database schema, inserting records, and scanning directories for images and videos. When processing images, it tries OCR extraction via FacebookAdScanner.basic_ocr (from facebookAd.py) and saves the text in the Content_Info column. The script also displays a formatted table of current records. It is useful for testing and illustrating how OCR-derived text supports downstream matching in main.py and newsAPI.py.

```
"""
```

This section of code is intended to scan a Facebook user's domain for advertisements. If advertisements are detected, their details will be retrieved and stored in a database named 'fbContentType.db' for further analysis. Since this is a hypothetical example, the actual process of scanning Facebook for ads is not included. Instead, the code will search the local file system of this project for files named 'image_by_jon_smith.jpg' or 'video_by_jane_doe.mp4' to simulate the presence of ads.

```
"""
```

```
import sqlite3
import os
from pathlib import Path
from facebookAd import FacebookAdScanner
```

```
# Column widths for table display
```

```
# The widths are slightly adjusted here for better accommodation of headers
```

```
"""
```

The widths of the columns for the database table display are defined in the COLUMN_WIDTHS list.

This allows for better formatting of the output table, with each width corresponding to a specific column in the fbContentType database table.

```
"""
```

```
COLUMN_WIDTHS = [10, 20, 12, 12, 9, 9, 9, 12, 40]
```

```
# Directory to scan for image and video files
```

```
"""
```

The SCAN_DIRECTORY constant specifies the directory to scan for image and video files.

For this coding project, the directory is named 'image_and_video_directory'.

This directory is stored locally within the project structure.

.....

```
SCAN_DIRECTORY = 'image_and_video_directory'
```

Placeholder sample data - for demonstration and testing purposes

.....

This list contains sample records to be inserted into the fbContentType database table.

Each entry represents a record with fields such as file type, file name, ad flags, engagement metrics, location, and content information.

.....

.....

The database structure is as follows, which can be seen in the def create_facebook_ads_table() function:

- File_Type: Type of the file (image or video)
- File_Name: Name of the file
- Facebook_Ad: Boolean flag indicating if it's a Facebook ad
- User_Content: Boolean flag indicating if it's user-generated content
- No_Shares: Number of shares
- No_Comments: Number of comments
- No_Likes: Number of likes
- Locations: Location information
- Content_Info: Extracted content information from OCR or metadata

.....

```
PLACEHOLDER_DATA = [
```

(

```
    'image',  
    'hello_kitty.jpg',  
    1, 0, 500, 85, 1200,  
    'London, UK',  
    'Hello Kitty is come to town!, Book now for your exclusive offers.'
```

),

(

```
    'video',  
    'hello_kitty_video.mp4',  
    1, 0, 500, 85, 1200,  
    'London, UK',  
    'Hello Kitty - LIVE!, Book now for your exclusive offer, limited time only.'
```

),

(

```

        'image',
        'david_image_labour.png',
        0, 1, 10, 5, 0,
        'Manchester, UK',
        'Labour Government does not like cats. Evidence shows that David is a Parrot lover.'
    )
]

```

""" This function establishes a connection to the SQLite database specified by db_path. """

```

def get_db_connection(db_path=':memory:'):
    """Establishes a connection to the SQLite database specified by db_path."""
    return sqlite3.connect(db_path)

```

""" This function creates the fbContentType table in the database if it does not already exist. """

```

def create_facebook_ads_table(cursor):
    """Creates the fbContentType table in the database if it does not already exist."""
    cursor.execute("""
        CREATE TABLE IF NOT EXISTS fbContentType (
            File_Type TEXT NOT NULL,
            File_Name TEXT NOT NULL,
            Facebook_Ad BOOLEAN DEFAULT 0,
            User_Content BOOLEAN DEFAULT 0,
            No_Shares INTEGER DEFAULT 0,
            No_Comments INTEGER DEFAULT 0,
            No_Likes INTEGER DEFAULT 0,
            Locations TEXT,
            Content_Info TEXT,
            Created_At TIMESTAMP DEFAULT CURRENT_TIMESTAMP
        )
    """)

```

```
def clear_table(cursor):
```

```

    """ Clears all records from the fbContentType table. """
    cursor.execute('DELETE FROM fbContentType')

```

""" This function inserts multiple records into the fbContentType table using executemany for efficiency. """

```
def insert_multiple_fb_content_types(cursor, content_data_list):
```

```

""" Inserts multiple records into the fbContentType table using executemany for efficiency. """
cursor.executemany("

    INSERT INTO fbContentType (
        File_Type, File_Name, Facebook_Ad, User_Content,
        No_Shares, No_Comments, No_Likes, Locations, Content_Info
    ) VALUES (?, ?, ?, ?, ?, ?, ?, ?, ?)
", content_data_list)

"""

This function formats a row of data for table display, centering each item within its column width."""

def format_row(data, widths):

    """
    Format a row of data with centered alignment and fixed widths.
    Truncates long data and adds an ellipsis for readability.
    """

    is_header = all(isinstance(item, str) for item in data)
    formatted_data = []

    for item, width in zip(data, widths):
        item_str = str(item)
        # Truncate if not a header AND the string is longer than the column width
        display_str = item_str if (is_header or len(item_str) <= width) else item_str[:width - 3] + '...'
        formatted_data.append(f'{display_str:{width}}')

    return '|' + '|'.join(formatted_data) + '|'

"""

This function prints the contents of the fbContentType table in a formatted table layout. """

def print_table(cursor):

    """
    Print database contents in a formatted table.
    """

    cursor.execute('SELECT * FROM fbContentType')
    all_rows = cursor.fetchall()
    column_names = [description[0] for description in cursor.description]

    print("\n--- Database Content Table ---")
    header_row = format_row(column_names, COLUMN_WIDTHS)
    separator = '-' * len(header_row)

    print(f'{separator}\n{header_row}\n{separator}')
    for row in all_rows:

```

```

print(f"{{format_row(row, COLUMN_WIDTHS)}\n{separator}}")

def scan_files_for_ads(directory):
    """
    The program scans a specified directory for image and video files. It uses Optical Character Recognition (OCR) to extract text from these images and stores the extracted information in a variable called Content_Info. Additionally, it determines whether the files are advertisements based on filename patterns, specifically looking for the presence of the word "ad" or certain keywords. Finally, the program returns a list of tuples that are formatted for database insertion.

    """
    if not os.path.exists(directory):
        print(f"Warning: Directory '{directory}' not found. Creating it...")
        os.makedirs(directory, exist_ok=True)
    return []

# Initialize OCR scanner
scanner = FacebookAdScanner()

# Media file extensions
"""
The media_extensions dictionary defines the file extensions for images and videos. This helps in identifying the type of media file during the scanning process.

"""

media_extensions = {
    'image': {'.jpg', '.jpeg', '.png', '.gif', '.webp', '.bmp'},
    'video': {'.mp4', '.avi', '.mov', '.mkv', '.webm', '.flv'}
}

"""
The `content_data` list is initially empty and will store the tuples of data that need to be inserted into the database.

Each tuple corresponds to a media file found in the specified directory, along with its associated information.

The function processes each file in the directory, checking its extension to determine if it's an image or a video. For images, it performs Optical Character Recognition (OCR) to extract text. Additionally, it identifies whether the file is an advertisement based on its filename and adds the relevant data to the `content_data` list.

"""

content_data = []

```

```

for filename in os.listdir(directory):
    file_path = os.path.join(directory, filename)
    if os.path.isdir(file_path):
        continue

    file_ext = Path(filename).suffix.lower()

    # Determine file type
    file_type = None
    for media_type, extensions in media_extensions.items():
        if file_ext in extensions:
            file_type = media_type
            break

    if not file_type:
        continue

    # Extract OCR text from images
    content_info = f'File detected: {filename}'
    if file_type == 'image':
        try:
            ocr_text = scanner.basic_ocr(file_path)
            if ocr_text and ocr_text.strip():
                content_info = ocr_text.strip()
                print(f"  OCR extracted from {filename}: {content_info[:60]}...")
        except Exception as e:
            print(f"  Warning: Could not extract OCR from {filename}: {e}")

    # Determine if it's an ad
    is_ad = any(keyword in filename.lower() for keyword in ['ad', 'advertisement', 'sponsored'])

    content_data.append((
        file_type, filename,
        1 if is_ad else 0, 0 if is_ad else 1,
        0, 0, 0, 'Unknown', content_info
    ))

return content_data
"""

```

The final `main()` function showcases the operations performed on the database. It creates a table, inserts placeholder data, scans for media files, and prints the resulting table. Additionally, it includes an if-else statement to handle the scenario where no media files are found.

.....

```
def main():
    """Main function to demonstrate database operations."""
    with get_db_connection('fbContentType.db') as conn:
        cursor = conn.cursor()
        print("Database connection opened and cursor created.")

        create_facebook_ads_table(cursor)
        print("Table 'fbContentType' created successfully.")

        clear_table(cursor)
        print("Cleared existing records from table.")

        # Insert placeholder data
        print(f"\nInserting placeholder sample data...")
        insert_multiple_fb_content_types(cursor, PLACEHOLDER_DATA)
        print(f"Successfully inserted {len(PLACEHOLDER_DATA)} placeholder records.")

        # Scan and insert files
        print(f"\nScanning '{SCAN_DIRECTORY}' for media files...")
        scanned_data = scan_files_for_ads(SCAN_DIRECTORY)

        if scanned_data:
            insert_multiple_fb_content_types(cursor, scanned_data)
            print(f"Successfully inserted {len(scanned_data)} scanned file records.")
        else:
            print(f"No media files found in '{SCAN_DIRECTORY}'.")

        print_table(cursor)

    print("\nDatabase connection closed and changes committed automatically.")
    print("Data saved to 'fbContentType.db' file.")
```

.....

The `if __name__ == '__main__':` block ensures the `main()` function runs only when the script is executed directly,

```
not when imported as a module. This practice is common in Python programming, allowing users to see how the script works.
```

```
"""
```

```
if __name__ == '__main__':
    main()
```

facebookAd.py

The FacebookAdScanner class offers OCR functionality, including a basic OCR method and an improved preprocessing pipeline that involves resizing, converting to grayscale, applying CLAHE, and thresholding to recover faint text using Tesseract with Pillow/OpenCV. When executed directly, it retrieves filenames from the database, resolves them against image_and_video_directory/, runs OCR on available images, and heuristically detects ad-related keywords like “Sponsored”. This module is utilised by database.py (to extract text for storage) and app/ocr.py (to enhance items dynamically), forming the core OCR component of the project’s verification process flow.

```
"""
```

```
This section of code is designed to scan a Facebook user domain to identify any advertisements. If advertisements are found, their details will be retrieved and stored in a database called 'fbContentType.db' for further analysis. Since this is a hypothetical example, the actual process of scanning Facebook for ads is not included. Instead, the code will search the local file system of this project for demo media to simulate the presence of ads and will run optical character recognition (OCR) on the images using Tesseract, if available.
```

```
Please note that `pytesseract` is a wrapper for the native `tesseract` binary. It is necessary to have the `tesseract` executable installed and available in your system's PATH for OCR to function properly.
```

```
"""
```

```
""" Import necessary libraries to handle file operations and database interactions. """
```

```
import os
import cv2
from PIL import Image
import pytesseract
```

```
"""
```

```
This class will initially scan for advertisements within the Facebook user domain. If any ads are found, the details will be stored in a database. Since this is a hypothetical example, the actual scanning process has not been implemented.
```

```

Instead, the class will search for specific files in the local file system located in the 'image_and_video_directory/'.
The information about each advertisement will be saved in the 'fbContentType.db' database.

#####
class FacebookAdScanner:
    def __init__(self, ad_keywords=None):
        #####
        First we need to check for 'Advertisement' or 'Sponsored' keywords in the text extracted from images.
        This will identify if the content is an ad.
        #####
        if ad_keywords is None:
            ad_keywords = ["Advertisement", "Sponsored"]
        self.ad_keywords = ad_keywords

    def contains_ad_keywords(self, text):
        #####
        Checks if the text contains advertising-related keywords.
        Returns (True, matched_keyword) if any keyword is found, False otherwise.
        #####
        text_lower = text.lower()
        for keyword in self.ad_keywords:
            keyword_lower = keyword.lower()
            if keyword_lower in text_lower:
                return (True, keyword)
        return False

    def ocr_with_preprocessing(self, image_path, upscale_fx=2, upscale_fy=2):
        #####
        Performs OCR with preprocessing for enhanced faint text recovery.
        Returns a list with results from both normal and inverted contrast.
        #####
        img = cv2.imread(image_path)
        if img is None:
            raise FileNotFoundError(f"Could not open or find image: {image_path}")

        # Upscale to improve OCR accuracy for faint text
        img = cv2.resize(img, (0, 0), fx=upscale_fx, fy=upscale_fy)
        gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)

        results = []
        for processed in [gray, cv2.bitwise_not(gray)]:

```

```

# Enhance contrast using CLAHE
clahe = cv2.createCLAHE(clipLimit=4.0, tileGridSize=(8,8))
enhanced = clahe.apply(processed)

# Adaptive thresholding
adaptive = cv2.adaptiveThreshold(
    enhanced, 255, cv2.ADAPTIVE_THRESH_MEAN_C, cv2.THRESH_BINARY, 15, 10)
text = pytesseract.image_to_string(Image.fromarray(adaptive))
results.append(text)

return results

""" This function performs the basic OCR on the original image without preprocessing. """
def basic_ocr(self, filepath):
    image = Image.open(filepath)
    return pytesseract.image_to_string(image)

""" If run as main, demonstrate the OCR and ad keyword detection. """
if __name__ == "__main__":
    scanner = FacebookAdScanner()
    filepath = "image_and_video_directory/hq720.jpg"
    ocr_results = scanner.ocr_with_preprocessing(filepath)
    for i, text in enumerate(ocr_results):
        print(f"----- Preprocessing Version {i+1} -----\\n{text}")
        result = scanner.contains_ad_keywords(text)
        print(f"Contains ad keywords: {result}\\n")
        if result and result[0]:
            print("Ad keyword detected: Needs to be verified!\\n")
    # Vanilla OCR baseline
    basic_text = scanner.basic_ocr(filepath)
    print("----- Basic OCR result -----")
    print(basic_text)
    print(f"Contains ad keywords: {scanner.contains_ad_keywords(basic_text)}")

```

newsAPI.py

This standalone, comprehensive script fetches articles from multiple news sources such as NewsAPI.org, NewsData.io, WorldNewsAPI, and optionally TheNewsAPI. It combines results, calculates basic text similarity using difflib.SequenceMatcher between OCR/database content and article titles or descriptions, and outputs relevant articles with their URLs. The

script also provides NEWS_API_SOURCES, which stores each provider's configuration details, including API keys and whether they are enabled. This setup is imported by the modular app/news.py to serve as a unified reference. Use this script for demonstrations, diagnostics, or testing queries and matches outside the main pipeline in main.py.

```
"""
```

```
This script uses the News API library to retrieve news articles and compares them with data from the  
fbContentType.db database.
```

```
It searches for similarities between Facebook ad content and news stories to help identify potential  
misinformation or related reports.
```

```
"""
```

```
""" Import necessary libraries and modules """
```

```
import sqlite3  
import os  
import requests  
from newsapi import NewsApiClient  
from difflib import SequenceMatcher  
from facebookAd import FacebookAdScanner
```

```
# Database configuration
```

```
DB_NAME = 'fbContentType.db'
```

```
"""
```

```
The NEW_API_SOURCES dictionary outlines various news API sources along with their corresponding API keys,  
allowing for seamless switching and fallback between different providers.
```

```
"""
```

```
# Multiple News API sources configuration
```

```
# Configure your API keys below (get free keys from respective websites)
```

```
NEWS_API_SOURCES = {
```

```
    'newsapi_org': {
```

```
        'api_key': '7db691a7480b4488b8c544b417996e8a', # NewsAPI.org (100 calls/day, 20 articles/call)
```

```
        'enabled': True,
```

```
        'url': 'https://newsapi.org'
```

```
    },
```

```
    'newsdata_io': {
```

```
        'api_key': 'pub_b67076ff2fb54340ae86b52ef2a65298', # NewsData.io (500 calls/month, 10 articles/call)
```

```
        'enabled': True,
```

```
        'url': 'https://newsdata.io'
```

```
    },
```

```

'thenewsapi': {
    'api_key': 'YOUR_THENEWSAPI_KEY', # TheNewsAPI.com (1M articles/week)
    'enabled': False,
    'url': 'https://thenewsapi.com'
},
'worldnewsapi': {
    'api_key': 'b145d010d6ea49f796b95b719833d013', # WorldNewsAPI.com (semantic tagging, sentiment)
    'enabled': True,
    'url': 'https://worldnewsapi.com'
}
}

def get_db_connection(db_path=DB_NAME):
    """ Establishes a connection to the SQLite database. """
    return sqlite3.connect(db_path)

def fetch_content_from_database():
    """
    Retrieves all content from the fbContentType database and returns a list of
    dictionaries containing file information and content.
    """

    with get_db_connection() as conn:
        cursor = conn.cursor()
        cursor.execute("""
            SELECT File_Name, Content_Info, Locations, Facebook_Ad, User_Content
            FROM fbContentType
        """)
        rows = cursor.fetchall()
        content_list = []
        scanner = FacebookAdScanner()
        for row in rows:
            file_name = row[0]
            # Try OCR if file exists
            ocr_text = None
            try:
                if file_name and os.path.exists(file_name):
                    ocr_text = scanner.basic_ocr(file_name)
            except Exception:

```

```

    ocr_text = None
    content_list.append({
        'file_name': file_name,
        'content_info': row[1],
        'ocr_text': ocr_text,
        'location': row[2],
        'is_facebook_ad': bool(row[3]),
        'is_user_content': bool(row[4])
    })
return content_list

"""

The fetch_media_content function scans a directory for media files,
extracts text from images using OCR, and identifies potential Facebook ads based on their filenames.

"""

def fetch_from_newsapi_org(query, api_key, language='en', page_size=20):
    """Fetch from NewsAPI.org"""
    try:
        newsapi = NewsApiClient(api_key=api_key)
        response = newsapi.get_everything(
            q=query,
            language=language,
            page_size=page_size,
            sort_by='relevancy'
        )
        return response.get('articles', [])
    except Exception as e:
        print(f"⚠️ NewsAPI.org error: {e}")
        return []

"""

This function named fetch_from_newsdata_io fetches articles from NewsData.io API.
"""

def fetch_from_newsdata_io(query, api_key, language='en', page_size=10):
    """ Fetch from NewsData.io """
    try:
        url =
f"https://newsdata.io/api/1/news?apikey={api_key}&q={query}&language={language}&page_size={page_size}"
        response = requests.get(url, timeout=10)
        data = response.json()
    
```

```

articles = []
for item in data.get('results', []):
    articles.append({
        'title': item.get('title', ''),
        'description': item.get('description', ''),
        'source': {'name': item.get('source_id', 'Unknown')},
        'url': item.get('link', ''),
        'publishedAt': item.get('pubDate', '')
    })
return articles
except Exception as e:
    print(f"⚠️ NewsData.io error: {e}")
    return []

```

""" This function named fetch_from_thenewsapi fetches articles from TheNewsAPI.com API and also handles errors. """

```

def fetch_from_thenewsapi(query, api_key, language='en', page_size=20):
    """Fetch from TheNewsAPI.com"""

    try:
        url =
f"https://api.thenewsapi.com/v1/news/all?api_token={api_key}&search={query}&language={language}&limit={page_size}"
        response = requests.get(url, timeout=10)
        data = response.json()

        articles = []
        for item in data.get('data', []):
            articles.append({
                'title': item.get('title', ''),
                'description': item.get('description', ''),
                'source': {'name': item.get('source', 'Unknown')},
                'url': item.get('url', ''),
                'publishedAt': item.get('published_at', '')
            })
        return articles
    except Exception as e:
        print(f"⚠️ TheNewsAPI.com error: {e}")
        return []

```

```

""" The newsAPI.py function named fetch_from_worldnewsapi fetches articles from WorldNewsAPI.com API. """
def fetch_from_worldnewsapi(query, api_key, language='en', page_size=20):
    """Fetch from WorldNewsAPI.com"""

    try:
        url = f"https://api.worldnewsapi.com/search-news?api-
key={api_key}&text={query}&language={language}&number={page_size}"
        response = requests.get(url, timeout=10)
        data = response.json()

        articles = []
        for item in data.get('news', []):
            articles.append({
                'title': item.get('title', ''),
                'description': item.get('text', '')[:200],
                'source': {'name': item.get('source_country', 'Unknown')},
                'url': item.get('url', ''),
                'publishedAt': item.get('publish_date', '')
            })
        return articles
    except Exception as e:
        print(f"⚠️ WorldNewsAPI.com error: {e}")
        return []

```

""" This function named fetch_news_articles fetches articles from multiple news API sources with fallback. """

```
def fetch_news_articles(query, language='en', page_size=20):
    """
    Fetches news articles from multiple sources with fallback.

```

Parameters:

- query (str): The search query (keywords to search for).
- language (str): Language of the articles (default: 'en').
- page_size (int): Number of articles to fetch (default: 20).

Returns:

- list: List of article dictionaries from all enabled sources.

"""

```
all_articles = []

```

```

sources_tried = []

# Try each enabled source
"""

The various if statements below check if each news API source is enabled. The if enabled block
fetches articles from that source and appends them to the all_articles list.

"""

if NEWS_API_SOURCES['newsapi_org']['enabled']:
    sources_tried.append('NewsAPI.org')
    articles = fetch_from_newsapi_org(
        query,
        NEWS_API_SOURCES['newsapi_org']['api_key'],
        language,
        page_size
    )
    all_articles.extend(articles)

if NEWS_API_SOURCES['newsdata_io']['enabled']:
    sources_tried.append('NewsData.io')
    articles = fetch_from_newsdata_io(
        query,
        NEWS_API_SOURCES['newsdata_io']['api_key'],
        language,
        min(page_size, 10)
    )
    all_articles.extend(articles)

if NEWS_API_SOURCES['thenewsapi']['enabled']:
    sources_tried.append('TheNewsAPI.com')
    articles = fetch_from_thenewsapi(
        query,
        NEWS_API_SOURCES['thenewsapi']['api_key'],
        language,
        page_size
    )
    all_articles.extend(articles)

if NEWS_API_SOURCES['worldnewsapi']['enabled']:
    sources_tried.append('WorldNewsAPI.com')
    articles = fetch_from_worldnewsapi(

```

```

query,
NEWS_API_SOURCES['worldnewsapi']['api_key'],
language,
page_size
)
all_articles.extend(articles)

if sources_tried:
    print(f" Searched: {', '.join(sources_tried)})")

return all_articles

```

def calculate_similarity(text1, text2):

"""

Calculates similarity ratio between two texts using SequenceMatcher.

Parameters:

- text1 (str): First text string.
- text2 (str): Second text string.

Returns:

- float: Similarity ratio (0.0 to 1.0).

"""

```

return SequenceMatcher(None, text1.lower(), text2.lower()).ratio()

```

def extract_keywords(content_info):

"""

Extracts potential keywords from content info for News API search.

Parameters:

- content_info (str): Content information text.

Returns:

- list: List of keyword strings.

"""

```

# Remove common words and extract meaningful keywords
"""

```

The code includes a list of common words to filter from the content_info.

```
This is done to concentrate on more significant words that are likely to  
produce better search results when querying news articles.
```

```
"""
```

```
common_words = {'the', 'a', 'an', 'and', 'or', 'but', 'in', 'on', 'at', 'to', 'for',  
    'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had'}  
  
words = content_info.lower().replace('\'', "").replace('.', "").replace("!", "").split()  
keywords = [word for word in words if word not in common_words and len(word) > 3]  
  
return keywords[:5] # Return top 5 keywords
```

```
"""
```

```
The function named cross_match_content_with_news matches database content with  
news articles from News API based on a specified similarity threshold.
```

```
"""
```

```
def cross_match_content_with_news(db_content_list, similarity_threshold=0.3):
```

```
"""
```

```
Cross-matches database content with news articles from News API.
```

Parameters:

```
NEWS_API_KEY = 'YOUR_API_KEY_HERE'      NEWS_API_KEY = 'YOUR_API_KEY_HERE'
```

```
db_content_list (list): List of content dictionaries from database.
```

```
similarity_threshold (float): Minimum similarity ratio to consider a match (0.0-1.0).
```

Returns:

```
list: List of matches with similarity scores.
```

```
"""
```

```
""" Initialize empty list to hold matches. """
```

```
matches = []
```

```
"""
```

```
The for loop below iterates through each content item in the database's content list.
```

```
For each item, it extracts keywords, constructs a search query, fetches related news articles, and  
compares the content with each article to calculate their similarity.
```

```
If the similarity exceeds the defined threshold, a match is recorded.
```

```
"""
```

```
for content in db_content_list:
```

```

print(f"\n{'='*80}")

print("Analyzing: {content['file_name']}")

print("Content: {content['content_info']}")

if content.get('ocr_text'):

    print("OCR Text: {content['ocr_text'][:100]}")

print("Type: {'Facebook Ad' if content['is_facebook_ad'] else 'User Content'}")

# Use OCR text if available, else content_info

search_text = content.get('ocr_text') or content['content_info']

keywords = extract_keywords(search_text)

query = ' '.join(keywords)

print("Search keywords: {query}")

# Fetch related news articles

articles = fetch_news_articles(query)

```

.....

An if block checks if any articles were returned from the fetch_news_articles function.

If no articles are found, it prints a message and continues to the next content item.

.....

```
if not articles:
```

```
    print("No news articles found for this content.")

    continue
```

```
print(f"Found {len(articles)} news articles. Analyzing similarities...")
```

Compare with each article

```
for article in articles:
```

```
    article_title = article.get('title', "")

    article_description = article.get('description', "")

    article_text = f"{article_title} {article_description}"
```

Calculate similarity

```
similarity = calculate_similarity(search_text, article_text)
```

```
if similarity >= similarity_threshold:
```

```
    match = {
```

```
        'db_content': content,
```

```
        'news_article': {
```

```
            'title': article_title,
```

```
            'description': article_description,
```

```
            'source': article.get('source', {}).get('name', 'Unknown'),
```

```
            'url': article.get('url', ""),
```

```
            'published_at': article.get('publishedAt', "")
```

```
        },
```

```
        'similarity_score': similarity
```

```

        }

        matches.append(match)

        print(f" ✓ Match found! Similarity: {similarity:.2%}")

        print(f" Article: {article_title[:60]}...")

    return matches
}

""" Based on the cross-matched results, this function displays the matches in a formatted manner. """
def display_matches(matches):
    """
    Displays the matched content and news articles in a formatted way.

    Parameters:
        matches (list): List of match dictionaries.

    """

    if not matches:
        print("\n" + "="*80)
        print("No matches found between database content and news articles.")
        return

    print("\n" + "="*80)
    print(f"SUMMARY: Found {len(matches)} match(es)")
    print("="*80)

    for i, match in enumerate(matches, 1):
        print(f"\n--- Match #{i} ---")
        print(f"Database File: {match['db_content']['file_name']}")
        print(f"DB Content: {match['db_content']['content_info']}")
        print(f"Type: {'Facebook Ad' if match['db_content'][is_facebook_ad] else 'User Content'}")
        print(f"\nMatched News Article:")
        print(f" Title: {match['news_article']['title']}")
        print(f" Source: {match['news_article']['source']}")
        print(f" Published: {match['news_article']['published_at']}")
        print(f" URL: {match['news_article']['url']}")
        print(f" Similarity Score: {match['similarity_score']:.2%}")
        print("-" * 80)

```

```

"""
The main() function executes the news API cross-matching process. This also allows users to see
how the script works and what results it produces.

"""

def main():
    """ Main function to run the news API cross-matching process. """
    print("*"*80)
    print("NEWS API CONTENT MATCHER")
    print("*"*80)
    print("This script compares content from fbContentType.db with news articles.\n")

    # Check if any API sources are enabled
    enabled_sources = [name for name, config in NEWS_API_SOURCES.items() if config['enabled']]

    if not enabled_sources:
        print("⚠ WARNING: No News API sources enabled!")
        print("Please enable at least one API source and add your API key.")
        print("\nAvailable sources:")
        print(" 1. NewsAPI.org - https://newsapi.org (100 calls/day)")
        print(" 2. NewsData.io - https://newsdata.io (500 calls/month)")
        print(" 3. TheNewsAPI.com - https://thenewsapi.com")
        print(" 4. WorldNewsAPI.com - https://worldnewsapi.com")
        print("\nRunning in demo mode with database content only...\n")

    # Just display database content
    content_list = fetch_content_from_database()
    print(f"Found {len(content_list)} items in database:")
    for item in content_list:
        print(f" - {item['file_name']}: {item['content_info'][:50]}...")
    return

    # Fetch content from database
    print("Step 1: Fetching content from database...")
    content_list = fetch_content_from_database()
    print(f"✓ Found {len(content_list)} items in database.\n")

    # Cross-match with news articles
    print("Step 2: Cross-matching with News API...")
    matches = cross_match_content_with_news(content_list, similarity_threshold=0.25)

```

```

# Display results
print("\n" + "="*80)
print("Step 3: Displaying Results")
display_matches(matches)

print("\n" + "="*80)
print("Analysis complete!")
print("="*80)

"""

The `if __name__ == '__main__':` block ensures that the `main()` function is called only when the script is executed directly, not when it is imported as a module into another script. This allows users to understand how the script works and the results it produces. This practice is common in Python programming.

"""

if __name__ == '__main__':
    main()

```

app/config.py

The central setup for the modular pipeline includes key settings like the database location, scan folder, similarity threshold, language, and page size. These settings are imported by other app modules to maintain consistency. The similarity threshold controls how strictly matches are accepted, while language and page size help shape the news API queries. Although API keys are stored separately, it's recommended to centralise all configurations, including secrets, using environment variables in this module for easier management. Any changes here will immediately impact the app.db, app.news, and app.match components, which rely on these settings main.py.

"""

The app/config.py module outlines the application's configuration constants, including the database name, scan directory, news API key, similarity threshold, page size, and language settings. These constants are used consistently throughout the application to maintain uniform configuration values.

The variables defined in this module include:

- DB_NAME: The name of the SQLite database file.
- SCAN_DIRECTORY: The directory path where images and videos are scanned.
- NEWS_API_KEY: The API key used for accessing news services.
- SIMILARITY_THRESHOLD: The threshold value for determining similarity in matching.

```
- PAGE_SIZE: The number of items to fetch per page from news APIs.  
- LANGUAGE: The language setting for content retrieval.
```

```
"""
```

```
DB_NAME = 'fbContentType.db'  
SCAN_DIRECTORY = 'image_and_video_directory'  
NEWS_API_KEY = '7db691a7480b4488b8c544b417996e8a'  
SIMILARITY_THRESHOLD = 0.15  
PAGE_SIZE = 20  
LANGUAGE = 'en'
```

app/db.py

The data access layer in the modular architecture manages connections to the SQLite database identified by app.config.DB_NAME. It transforms rows from fbContentType into ContentItem objects stored in app.models. This layer offers fetch_content_items, which supplies data to main.py, and clear_table for maintenance or testing. Centralising SQL operations here helps keep other modules database-agnostic. Additionally, this file is the ideal place to add indices, migrations, or extra queries- such as filtering by ad flag or time windows- without needing to change the main pipeline code.

```
"""
```

The `app/db.py` module contains functions for interacting with an SQLite database, specifically for retrieving content items and clearing the content table. It starts by importing the `sqlite3` library, which is used for database operations, as well as the `ContentItem` model that represents individual content items.

This module defines functions to establish a database connection, fetch content items from the `fbContentType` table, and delete all entries from that table. Each function is designed to perform database operations in a clear and straightforward manner, allowing the application to efficiently access and manage the content item data stored

in the SQLite database.

```
"""
```

```
import sqlite3  
from typing import List  
from .models import ContentItem  
from .config import DB_NAME
```

```
"""
```

The function `get_db_connection` creates a connection to the SQLite database using the specified database path, defaulting to `DB_NAME` as indicated in the configuration. It returns a connection object for database interaction.

.....

```
def get_db_connection(db_path: str = DB_NAME):  
    return sqlite3.connect(db_path)
```

.....

The `'fetch_content_items'` function retrieves all content items from the `'fbContentType'` table in the SQLite database.

It establishes a connection to the database, executes a SQL query to select the relevant fields, and then creates a

list of `'ContentItem'` objects based on the retrieved rows.

.....

```
def fetch_content_items() -> List[ContentItem]:  
    with get_db_connection() as conn:  
        c = conn.cursor()  
        c.execute(  
            """  
            SELECT File_Name, Content_Info, Locations, Facebook_Ad, User_Content  
            FROM fbContentType  
            """  
        )  
        rows = c.fetchall()  
        items: List[ContentItem] = []  
        for file_name, content_info, location, facebook_ad, user_content in rows:  
            items.append(  
                ContentItem(  
                    file_name=file_name,  
                    content_info=content_info or "",  
                    location=location or "",  
                    is_facebook_ad=bool(facebook_ad),  
                    is_user_content=bool(user_content),  
                )  
            )  
    return items
```

.....

The `clear_table` function removes all entries from the `fbContentType` table in the SQLite database.

It creates a connection, executes a SQL DELETE command, and commits the changes.

```
"""
def clear_table():
    with get_db_connection() as conn:
        c = conn.cursor()
        c.execute("DELETE FROM fbContentType")
        conn.commit()
```

app/models.py

Defines lightweight data models using dataclasses: ContentItem (a database row with optional ocr_text), NewsArticle (an abstract view of article fields), and MatchResult (associating an item with an article and a similarity score). These models enable consistent data transfer across modules (app.db, app.ocr, app.news, app.match, and main.py). By centralising this structure, refactoring, typing, and testing are simplified, while keeping presentation or provider-specific details separate from core logic. For stricter validation, upgrading to models like Pydantic is recommended if stricter validation is necessary.

"""

The app/models.py module defines the data models used in the application, which include ContentItem, NewsArticle, and MatchResult. These models are structured as dataclasses to simplify the storage and manipulation of related data attributes.

"""

```
from dataclasses import dataclass
from typing import Optional
```

"""

The @dataclass decorator is used to automatically generate special methods for the class, such as __init__() and __repr__(), based on the defined attributes.

The class ContentItem defined below represent the core data structures used throughout the application:

- ContentItem: Represents a content item with attributes such as file_name, content_info, location, is_facebook_ad, is_user_content, and an optional ocr_text for storing OCR results.
- NewsArticle: Represents a news article with attributes like title, description, source, url, and published_at.
- MatchResult: Represents the result of matching a ContentItem with a NewsArticle, including the similarity score.

"""

```
@dataclass
class ContentItem:
```

```
file_name: str
content_info: str
location: str
is_facebook_ad: bool
is_user_content: bool
ocr_text: Optional[str] = None
```

.....

The `NewsArticle` class represents a news article and includes attributes such as title, description, source, URL, and publication date.

It is implemented as a dataclass, using the `@dataclass` decorator to automatically generate special methods for the class, including

`__init__()` and `__repr__()`, based on the defined attributes.

.....

```
@dataclass
class NewsArticle:
    title: str
    description: str
    source: str
    url: str
    published_at: str
```

.....

The final class, `MatchResult`, represents the outcome of matching a `ContentItem` with a `NewsArticle`, including a similarity score.

This class is implemented as a data class. The `@dataclass` decorator is utilized to automatically generate special methods for the

class, such as `__init__()` and `__repr__()`, based on the specified attributes.

.....

```
@dataclass
class MatchResult:
    item: ContentItem
    article: NewsArticle
    similarity: float
```

app/ocr.py

The runtime OCR enrichment in the modular pipeline processes a list of ContentItems by attempting to read each file path and, if available, uses FacebookAdScanner.basic_ocr (from facebookAd.py) to fill in ocr_text. This straightforward function allows main.py to treat OCR as an optional step, enabling it to continue even if files are missing. It complements database.py's ingestion-time OCR by updating or completing data during analysis without changing the database contents.

.....

This module provides functionality to enhance content items by extracting text from image files using optical character recognition (OCR). It begins by importing the necessary libraries and the `ContentItem` model. The `os` module is used to check whether a file exists, and the `List` type from the `typing` module is employed for type hinting.

The `enrich_with_ocr` function processes a list of `ContentItem` objects. For each item, it checks if the associated file exists. If the file is present, the function utilizes the `FacebookAdScanner` class to perform OCR on the image file. The extracted text is then assigned to the `ocr_text` attribute of the `ContentItem`. If any exceptions occur during this process, the `ocr_text` attribute is set to `None`.

.....

```
import os
from typing import List
from .models import ContentItem
from facebookAd import FacebookAdScanner
```

.....

This function enhances a list of ContentItem objects by extracting OCR (Optical Character Recognition) text from their associated image files. It includes a conditional statement to check if the file exists before attempting to perform OCR. Additionally, a try-except block is used to handle any exceptions that may occur during the OCR process, ensuring that the program runs smoothly even if some files cannot be processed.

.....

```
def enrich_with_ocr(items: List[ContentItem]) -> List[ContentItem]:
    """ Enriches a list of ContentItem objects with OCR text extracted from their associated image files. """
    scanner = FacebookAdScanner()
    for item in items:
        try:
            if item.file_name and os.path.exists(item.file_name):
```

```
    item.ocr_text = scanner.basic_ocr(item.file_name)
except Exception:
    item.ocr_text = None
return items
```

app/news.py

A unified interface for querying multiple news providers has been implemented. It defines individual fetchers for providers such as NewsAPI.org, NewsData.io, WorldNewsAPI, and optionally TheNewsAPI, normalising various responses into NewsArticle objects and combining the results. The NEWS_API_SOURCES from newsAPI.py is imported for consistent configuration with the standalone script. The fetch_news_articles(query) function is passed to app.match.match_items_with_news by main.py, clearly separating data retrieval from the matching process. Consider adding caching, retries, or environment-based keys here to improve resilience and manage quotas efficiently.

```
"""
```

This script includes functions to fetch news articles from various news APIs based on a specified query. It starts by importing the necessary libraries and the NewsArticle model. The script defines separate functions to retrieve articles from NewsAPI.org, NewsData.io, TheNewsAPI, and WorldNewsAPI. Each function handles API requests, processes the responses, and returns a list of NewsArticle objects. The main function, `fetch_news_articles`, consolidates articles from all active sources based on the provided query.

```
"""
```

```
from typing import List
import requests
from newsapi import NewsApiClient
from .config import LANGUAGE, PAGE_SIZE
from .models import NewsArticle
```

```
# Import NEWS_API_SOURCES from the main newsAPI.py for now (could be moved to config)
```

```
"""
```

This code attempts to import `NEWS_API_SOURCES` from the `newsAPI` module if it is available. If the module is not present, it initializes `NEWS_API_SOURCES` as an empty dictionary. This fallback mechanism ensures that the script can still function without the `newsAPI` module. The import statement is enclosed in a try-except block to gracefully handle any potential `ImportError`.

```
"""
```

```
try:
    from newsAPI import NEWS_API_SOURCES
except ImportError:
```

```

NEWS_API_SOURCES = {}

"""

The fetch_from_newsapi_org function obtains news articles from NewsAPI.org using a specified query and API key.

It processes the API response and returns a list of NewsArticle objects.

"""

def fetch_from_newsapi_org(query, api_key, language=LANGUAGE, page_size=PAGE_SIZE):

    """ The block is wrapped in a try-except to handle any exceptions that may occur during the API request or processing. """

    try:
        newsapi = NewsApiClient(api_key=api_key)
        response = newsapi.get_everything(
            q=query,
            language=language,
            page_size=page_size,
            sort_by='relevancy'
        )
        articles = []
        for a in response.get('articles', []):
            articles.append(NewsArticle(
                title=a.get('title', ''),
                description=a.get('description', ''),
                source=(a.get('source') or {}).get('name', 'Unknown'),
                url=a.get('url', ''),
                published_at=a.get('publishedAt', '')
            ))
        print(f" ✓ NewsAPI.org: Found {len(articles)} articles")
        return articles
    except Exception as e:
        print(f" ⚠️ NewsAPI.org error: {e}")
        return []

"""

The function `fetch_from_newsdata_io` is designed to retrieve news articles from NewsData.io using a specified query and API key.

It processes the response from the API and returns a list of `NewsArticle` objects. To facilitate making HTTP requests, the function

```

imports the requests library. Additionally, a try-except block is implemented to handle any exceptions that may occur during the

API request or while processing the response.

=====

```
def fetch_from_newsdata_io(query, api_key, language=LANGUAGE, page_size=10):
    """ Fetches news articles from NewsData.io using the provided query and API key. """
    try:
        # NewsData.io uses 'size' parameter, not 'page_size'
        url =
f"https://newsdata.io/api/1/news?apikey={api_key}&q={query}&language={language}&size={page_size}"
        response = requests.get(url, timeout=10)
        data = response.json()

        # Check for API errors
        if data.get('status') == 'error':
            error_msg = data.get('results', {}).get('message', 'Unknown error') if isinstance(data.get('results'), dict) else
'Unknown error'
            print(f"⚠️ NewsData.io error: {error_msg}")
            return []

        articles = []
        for item in data.get('results', []):
            articles.append(NewsArticle(
                title=item.get('title', ""),
                description=item.get('description', ""),
                source=item.get('source_id', 'Unknown'),
                url=item.get('link', ""),
                published_at=item.get('pubDate', "")
            ))
        print(f"✓ NewsData.io: Found {len(articles)} articles")
        return articles
    except Exception as e:
        print(f"⚠️ NewsData.io error: {e}")
        return []

=====
```

This function processes the API response and returns a list of NewsArticle objects.

It also imports the requests library for making HTTP requests.

=====

```
def fetch_from_thenewsapi(query, api_key, language=LANGUAGE, page_size=PAGE_SIZE):
```

```

""" Fetches news articles from TheNewsAPI using the provided query and API key."""
try:
    url =
f"https://api.thenewsapi.com/v1/news/all?api_token={api_key}&search={query}&language={language}&limit={page_size}"
    response = requests.get(url, timeout=10)
    data = response.json()
    articles = []
    for item in data.get('data', []):
        articles.append(NewsArticle(
            title=item.get('title', ''),
            description=item.get('description', ''),
            source=item.get('source', 'Unknown'),
            url=item.get('url', ''),
            published_at=item.get('published_at', '')
        ))
    return articles
except Exception:
    return []

def fetch_from_worldnewsapi(query, api_key, language=LANGUAGE, page_size=PAGE_SIZE):
    """ Fetches news articles from WorldNewsAPI using the provided query and API key."""
    try:
        url = f"https://api.worldnewsapi.com/search-news?api-key={api_key}&text={query}&language={language}&number={page_size}"
        response = requests.get(url, timeout=10)
        data = response.json()
        articles = []
        for item in data.get('news', []):
            articles.append(NewsArticle(
                title=item.get('title', ''),
                description=item.get('text', '')[:200],
                source=item.get('source_country', 'Unknown'),
                url=item.get('url', ''),
                published_at=item.get('publish_date', '')
            ))
        print(f" ✓ WorldNewsAPI: Found {len(articles)} articles")
        return articles
    except Exception as e:

```

```

print(f"⚠️ WorldNewsAPI error: {e}")
return []

"""

The final function, fetch_news_articles, collects articles from all enabled sources based on the provided query. It includes various if statements to check which sources are active in NEWS_API_SOURCES.

"""

def fetch_news_articles(query: str) -> List[NewsArticle]:
    """ Fetches news articles from all enabled sources in NEWS_API_SOURCES."""
    all_articles = []
    if not NEWS_API_SOURCES:
        # fallback to NewsAPI.org with config key if NEWS_API_SOURCES not found
        from .config import NEWS_API_KEY
        return fetch_from_newsapi_org(query, NEWS_API_KEY)
    if NEWS_API_SOURCES.get('newsapi_org', {}).get('enabled'):
        all_articles.extend(fetch_from_newsapi_org(query, NEWS_API_SOURCES['newsapi_org']['api_key']))
    if NEWS_API_SOURCES.get('newsdata_io', {}).get('enabled'):
        all_articles.extend(fetch_from_newsdata_io(query, NEWS_API_SOURCES['newsdata_io']['api_key']))
    if NEWS_API_SOURCES.get('thenewsapi', {}).get('enabled'):
        all_articles.extend(fetch_from_thenewsapi(query, NEWS_API_SOURCES['thenewsapi']['api_key']))
    if NEWS_API_SOURCES.get('worldnewsapi', {}).get('enabled'):
        all_articles.extend(fetch_from_worldnewsapi(query, NEWS_API_SOURCES['worldnewsapi']['api_key']))
    return all_articles

```

app/match.py

Performs keyword extraction and similarity scoring to link database or OCR text with potential news articles. The extract_keywords function generates succinct queries by removing common stop words. The similarity function employs SequenceMatcher to evaluate text similarities, while match_items_with_news queries the news API and filters results based on the set similarity threshold (app.config.SIMILARITY_THRESHOLD). It produces a list of MatchResult objects used by main.py for reporting purposes. Since the inputs and outputs are typed (app.models), this module is straightforward to test and can be upgraded to incorporate more advanced, semantic matching techniques in the future work.

```

"""

The `app/match.py` module provides functionality for matching content items with news articles based on their textual similarity.

```

It begins by importing the necessary libraries and the data models used in the application.

The module defines functions to extract keywords from text, calculate similarity scores, and match content items with news articles from various sources. The matching process involves querying news APIs using the keywords extracted from the content items and comparing the text of these items with the titles and descriptions of the fetched news articles. If the similarity score exceeds a defined threshold, a match is recorded.

Additionally, the script imports the `SequenceMatcher` class from the `difflib` module to compute similarity ratios between strings, as well as the `List` type from the `typing` module for type hinting.

```
from difflib import SequenceMatcher
from typing import List
from .models import ContentItem, NewsArticle, MatchResult
from .config import SIMILARITY_THRESHOLD
```

The extract_keywords function identifies significant keywords from a text by removing common words and limiting the number

of keywords returned. This function helps in forming effective search queries to find news articles relevant to the content.

```
def extract_keywords(text: str) -> str:
    """ Extracts significant keywords from text by removing common words. """
    common = {
        'the', 'a', 'an', 'and', 'or', 'but', 'in', 'on', 'at', 'to', 'for', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had'
    }
    words = text.lower().replace('.', ' ').replace('!', ' ').split()
    keywords = [w for w in words if w not in common and len(w) > 3]
    return ' '.join(keywords[:5])
```

The similarity function computes a similarity score between two input strings by utilizing the SequenceMatcher class from the difflib module. It returns a float value that indicates the similarity ratio between the two strings.

This score is useful for assessing how closely the content item matches a news article.

```

"""
def similarity(a: str, b: str) -> float:
    """ Computes a similarity score between two strings using SequenceMatcher. """
    return SequenceMatcher(None, (a or "").lower(), (b or "").lower()).ratio()

"""

The `match_items_with_news` function is designed to match a list of `ContentItem` objects with news articles gathered from a specified fetcher function. It iterates through each content item, extracts keywords to formulate a search query, and retrieves news articles related to that query. The function then compares the text of each content item with the titles and descriptions of the retrieved news articles, calculating a similarity score for each comparison. If the similarity score exceeds a predefined threshold, a `MatchResult` object is created and added to a results list. Ultimately, the function returns a list of `MatchResult` objects that represent the successful matches.

"""

def match_items_with_news(items: List[ContentItem], fetcher) -> List[MatchResult]:
    """ Matches content items with news articles using similarity scoring. """
    results: List[MatchResult] = []
    for item in items:
        # ONLY use Content_Info from database (or OCR text if available)
        # Never match on filename
        base_text = item.ocr_text or item.content_info
        if not base_text or not base_text.strip():
            continue

        # Extract keywords from content only
        query = extract_keywords(base_text)
        if not query:
            continue

        print(f"\n Searching for: '{query}' (from {item.file_name})")
        articles = fetcher(query)
        print(f" Retrieved {len(articles)} articles total")

        for art in articles:
            # Compare content with news article title and description only

```

```

score = similarity(base_text, f'{art.title} {art.description}')
if score >= SIMILARITY_THRESHOLD:
    results.append(MatchResult(item=item, article=art, similarity=score))
    print(f"✓ Match! Score: {score:.2%} - {art.title[:60]}")
return results

```

app/__init__.py

Marks the app directory as a Python package, enabling clean module imports like from app import db, news. Although currently empty, this file supports the modular architecture used by main.py. If you later include package-level helpers or versioning information, this is the right spot. Keeping __init__.py simple avoids hidden side effects during imports and facilitates easier unit testing of individual submodules.

```

=====
app package
=====

This package contains the core modules used by the Facebook Ad Content
Verification System. It is intentionally lightweight: no runtime logic is
executed at import time. The purpose of this file is to mark the directory as a
Python package and provide a concise overview of the available submodules.

```

Modules

```

-----
- config: Global configuration constants (e.g., DB_NAME, SCAN_DIRECTORY,
NEWS_API_KEY, SIMILARITY_THRESHOLD, PAGE_SIZE, LANGUAGE).
- db: Database helpers for SQLite (connections, queries, and simple CRUD).
- models: Data models used across the app (e.g., NewsArticle).
- ocr: OCR pipeline built on Tesseract with image pre-processing utilities.
- news: Multi-source news fetching and aggregation utilities integrating
NewsAPI.org, NewsData.io, TheNewsAPI, and WorldNewsAPI.
- match: Similarity and matching utilities for comparing extracted content
against news articles.

```

Typical imports

```

-----
- from app.news import fetch_news_articles
- from app.ocr import extract_text_from_image # if defined in ocr.py

```

```

- from app.match import ...           # matching helpers

Note
-----
Keep this file minimal to avoid circular imports and side effects. Prefer
importing directly from the specific submodule you need (e.g., app.news,
app.ocr) rather than relying on package-level re-exports.

"""
# Public submodules for static analysis and auto-completion tools
__all__ = [
    "config",
    "db",
    "models",
    "ocr",
    "news",
    "match",
]
# Optional, lightweight package metadata
__version__ = "0.1.0"

```

Prerequisites

Required Software

- **Python 3.8 or higher**
- **Tesseract OCR engine** - Must be installed locally on your system (not a Python package)
- **Virtual environment** (recommended)

Required API Keys

You **must register** for free API keys from at least one of the following news services:

- **NewsAPI.org:** <https://newsapi.org/register> (100 requests/day free tier)
- **NewsData.io:** <https://newsdata.io/register> (500 requests/month free tier)
- **WorldNewsAPI:** <https://worldnewsapi.com/register> (Free tier available)

Important: The system will not function without at least one valid API key configured.

Installing Tesseract OCR (Required)

Tesseract must be installed **locally on your machine** before running this project. It is **not** a Python package and cannot be installed via pip.

macOS:

```
brew install tesseract
```

Linux (Ubuntu/Debian):

```
sudo apt-get install tesseract-ocr
```

Windows:

1. Download the installer from: <https://github.com/UB-Mannheim/tesseract/wiki>
2. Run the installer and follow the setup wizard
3. Add Tesseract to your system PATH during installation

Verify Installation:

```
tesseract --version
```

If installed correctly, you should see the Tesseract version information.

Installation

1. **Clone or navigate to the project directory:**

```
cd "../Dissertation Code"
```

2. **Create and activate virtual environment:**

```
python3 -m venv .venv  
source .venv/bin/activate # On macOS/Linux  
# or  
.venv\Scripts\activate # On Windows
```

3. **Install required dependencies:**

```
pip install -r requirements.txt
```

4. **Register for News API keys (REQUIRED):**

See Prerequisites → Required API Keys for registration links and free-tier quotas. You must have at least one valid key enabled.

5. **Configure API keys in the project:**

Update your API keys in newsAPI.py (line 18-40):

```

NEWS_API_SOURCES = {
    'newsapi_org': {
        'api_key': 'YOUR_NEWSAPI_ORG_KEY', # Replace with your actual key
        'enabled': True, # Set to True to enable this source
    },
    'newsdata_io': {
        'api_key': 'YOUR_NEWSDATA_IO_KEY', # Replace with your actual key
        'enabled': True,
    },
    'worldnewsapi': {
        'api_key': 'YOUR_WORLDNEWSAPI_KEY', # Replace with your actual key
        'enabled': True,
    },
}

```

Note: At minimum, enable at least one API source with a valid key. The system aggregates result from all enabled sources.

Project Structure

```

.
├── app/          # Modular application package
│   ├── __init__.py
│   ├── config.py      # Configuration settings
│   ├── db.py         # Database operations
│   ├── models.py     # Data models (ContentItem, NewsArticle, MatchResult)
│   ├── ocr.py        # OCR processing
│   ├── news.py       # Multi-source news API integration
│   └── match.py      # Content matching logic
├── database.py    # Database setup and scanning script
├── facebookAd.py # Facebook ad scanner with OCR
├── newsAPI.py     # News API cross-matching script
└── main.py        # Main entry point (modular pipeline)
└── image_and_video_directory/ # Directory for media files to analyse
└── fbContentType.db # SQLite database (created automatically)
└── README.md      # This file

```

Usage

1. Initialise Database and Scan Files

First, populate the database with media files from the image_and_video_directory/:

```
python database.py
```

This will:

- Create the fbContentType.db SQLite database
- Scan all images and videos in image_and_video_directory/
- Extract text using OCR from images

- Store content information in the database

2. Run the Main Pipeline

Execute the complete verification pipeline:

```
python main.py
```

This will:

- Load all content items from the database
- Perform OCR on image files
- Search for matching news articles across all enabled APIs
- Display matches with similarity scores and URLs
- Flag unmatched content as potential misinformation

3. Run News API Matcher (Standalone)

For detailed analysis with the monolithic script:

```
python newsAPI.py
```

This provides more verbose output including:

- Detailed similarity scores
- All API sources searched
- Step-by-step matching process

4. Scan Individual Files (OCR Testing)

Test OCR on files in the database:

```
python facebookAd.py
```

This will:

- Read all files from the database
- Perform OCR with preprocessing
- Detect ad keywords ("Sponsored", "Advertisement")
- Display OCR results for each file

Understanding the Output

Matched Content

Match #1 | score=20.00%

- File: hq720.jpg
- Content: FOX 32 | MAJOR AWS OUTAGE TAKES WEBSITES, APPS OFFLINE
- Article: Amazon Says AWS Cloud Service is Back to Normal After Outage
- Source: deccanchronicle
- URL: <https://www.deccanchronicle.com/technology/amazon-says-aws...>

Interpretation: Content was verified against a legitimate news source with 20% similarity.

Unmatched Content

 File: hello_kitty.jpg

Content: Hello Kitty is come to town!, Book now for your exclusive offers.

Status: The information from this post cannot be verified ,

please be cautious of potential misinformation.

Interpretation: No matching news articles found - potential false information or promotional content.

Configuration

Similarity Threshold

Adjust in app/config.py:

```
SIMILARITY_THRESHOLD = 0.15 # Lower = more matches, higher = stricter matching
```

News API Sources

Enable/disable sources in newsAPI.py:

```
NEWS_API_SOURCES = {  
    'newsapi_org': {  
        'api_key': 'YOUR_KEY',  
        'enabled': True, # Set to False to disable  
    },  
    # ...  
}
```

Database Configuration

Change database name or scan directory in app/config.py:

```
DB_NAME = 'fbContentType.db'
```

```
SCAN_DIRECTORY = 'image_and_video_directory'
```

How It Works

1. Content Extraction

Image/Video → Tesseract OCR → Text Extraction → Database Storage

2. Keyword Extraction

Content Text → Remove Stop Words → Extract Top 5 Keywords → Search Query

3. News Matching

Search Query → Multi-API Fetch → Similarity Calculation → Match Results

4. Similarity Scoring

Uses SequenceMatcher from Python's difflib to calculate text similarity:

- 0.0 = No similarity
- 1.0 = Identical text
- Threshold: 0.15 (configurable)

Dependencies

Core dependencies (install via requirements.txt):

- newsapi-python - NewsAPI.org client
- requests - HTTP requests for other APIs
- pytesseract - Python wrapper for Tesseract OCR
- opencv-python - Image preprocessing
- Pillow - Image handling
- numpy - Numerical operations

Troubleshooting

Tesseract Not Found

Error: TesseractNotFoundError

Solution: Install Tesseract OCR and ensure it's in your PATH.

API Rate Limit Exceeded

 NewsAPI.org error: rateLimited

Solution: Wait 24 hours for quota reset or enable other API sources.

No Matches Found

Found 0 matches

Possible causes:

- Similarity threshold too high (try lowering to 0.10-0.15)
- API quota exhausted (check enabled sources)
- Content is truly unverified (potential misinformation)

File Not Found Errors

 File not found: image.jpg

Solution: Ensure files are in image_and_video_directory/ or database contains full paths.

14. Reference:

Bakir, V. (2020). Psychological Operations in Digital Political Campaigns: Assessing Cambridge Analytica's Psychographic Profiling and Targeting. *Frontiers in Communication*, [online] 5. doi: <https://doi.org/10.3389/fcomm.2020.00067> [Accessed 9 Oct. 2025].

BBC Bitesize (2024). *Timeline of how online misinformation fuelled UK riots - BBC Bitesize*. [online] BBC Bitesize. Available at: <https://www.bbc.co.uk/bitesize/articles/zshjs82> [Accessed 9 Oct. 2025].

BBC News (2018b). The global reach of Cambridge Analytica. *BBC News*. [online] 22 Mar. Available at: <https://www.bbc.com/news/world-43476762> [Accessed 14 Oct. 2025].

Benesch, S. (2021). *Nobody Can See Into Facebook*. [online] Available at: http://www.businessforum.com/Atlantic_10-30-2021.pdf [Accessed 11 Aug. 2025].

Byte Myke (2021). *SQLite beginner crash course in Visual Studio Code - 2022*. [online] YouTube. Available at: https://www.youtube.com/watch?v=IBgWKTaG_Bs [Accessed 19 Oct. 2025].

Dam, V.H., Hjordt, L.V., Cunha-Bang, S., Sestoft, D., Knudsen, G.M. and Stenbæk, D.S. (2021). Trait aggression is associated with five-factor personality traits in males. *Brain and Behavior*, [online] 11(7). doi: <https://doi.org/10.1002/bbr.3.2175> [Accessed 3 Oct. 2025].

Dixon, S.J. (2025c). *Most Popular Social Networks Worldwide as of February 2025, by Number of Monthly Active Users*. [online] Statista. Available at: <https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/> [Accessed 4 Aug. 2025].

free-news-api (2024). *GitHub - free-news-api/news-api: Top Free News API Comparison*. [online] GitHub. Available at: <https://github.com/free-news-api/news-api> [Accessed 18 Oct. 2025].

Fung, B. (2024). *UK riots show how social media can fuel real-life harm. It's only getting worse*. [online] CNN. Available at: <https://edition.cnn.com/2024/08/09/tech/uk-protests-social-media> [Accessed 9 Oct. 2025].

GeeksforGeeks (2024). *Introduction to Python Pytesseract Package*. [online] GeeksforGeeks. Available at: <https://www.geeksforgeeks.org/python/introduction-to-python-ptesseract-package/> [Accessed 18 Oct. 2025].

History.com Editors (2018). *The 2016 U.S. Presidential Election*. [online] HISTORY. Available at: <https://www.history.com/articles/us-presidential-election-2016> [Accessed 3 Sep. 2025].

Joseph, J., Sivaraman, J., Periyasamy, R. and Simi, V.R. (2017). An objective method to identify optimum clip-limit and histogram specification of contrast limited adaptive histogram equalization for MR images. *Biocybernetics and Biomedical Engineering*, 37(3), pp.489–497. doi: <https://doi.org/10.1016/j.bbe.2016.11.006> [Accessed 19 Oct. 2025].

Kiderlin, S. (2024). *Online disinformation sparked a wave of far-right violence in the UK — here's how*. [online] CNBC. Available at: <https://www.cnbc.com/2024/08/09/online-disinformation-sparked-a-wave-of-far-right-violence-in-the-uk.html> [Accessed 9 Oct. 2025].

Kite (2000). *Sqlite 3 Python Tutorial in 5 minutes - Creating Database, Tables and Querying [2020] *. [online] www.youtube.com. Available at: <https://www.youtube.com/watch?v=girsuXz0yA8> [Accessed 18 Oct. 2025].

Lee, M. (2022). *pytesseract: Python-tesseract is a python wrapper for Google's Tesseract-OCR*. [online] PyPI. Available at: <https://pypi.org/project/pytesseract/> [Accessed 18 Oct. 2025].

Legislation.Gov.UK (2023). *Online Safety Act 2023*. [online] Legislation.gov.uk. Available at: <https://www.legislation.gov.uk/ukpga/2023/50> [Accessed 9 Oct. 2025].

Lim, A. (2025). Big Five Personality Traits: The 5-Factor Model Of Personality. *Simply Psychology*. [online] Available at: <https://www.simplypsychology.org/big-five-personality.html> [Accessed 1 Oct. 2025].

mattlisiv (2018). *GitHub - mattlisiv/newsapi-python: A Python Client for News API*. [online] GitHub. Available at: <https://github.com/mattlisiv/newsapi-python> [Accessed 19 Oct. 2025].

Meta (2022). *Facebook Data policy*. [online] Facebook. Available at: <https://www.facebook.com/about/privacy/update/printable> [Accessed 31 Aug. 2025].

Mohamed, E. (2024). *Southport stabbing: What led to the spread of disinformation? * [online] Al Jazeera. Available at: <https://www.aljazeera.com/news/2024/8/2/southport-stabbing-what-led-to-the-spread-of-disinformation> [Accessed 9 Oct. 2025].

Netflix (2019). *The Great Hack | Netflix Official Site*. [online] www.netflix.com. Available at: <https://www.netflix.com/gb/title/80117542> [Accessed 8 Sep. 2025].

Newsapi.org (2019). *News API - A JSON API for live news and blog articles*. [online] Newsapi.org. Available at: <https://newsapi.org/> [Accessed 19 Oct. 2025].

NewsData (n.d.). *NewsData - News API to Search & Collect Worldwide News*. [online] Newsdata. Available at: <https://newsdata.io/> [Accessed 19 Oct. 2025].

NumPy (2022). *NumPy Documentation*. [online] numpy.org. Available at: <https://numpy.org/doc/> [Accessed 19 Oct. 2025].

Nyabola, N. (2019). *The spectre of Cambridge Analytica still haunts African elections*. [online] Al Jazeera. Available at: <https://www.aljazeera.com/opinions/2019/2/15/the-spectre-of-cambridge-analytica-still-haunts-african-elections> [Accessed 14 Oct. 2025].

Opencv.org (n.d.). *OpenCV documentation index*. [online] docs.opencv.org. Available at: <https://docs.opencv.org/> [Accessed 19 Oct. 2025].

Pillow.readthedocs.io (n.d.). *Pillow — Pillow (PIL Fork) 7.2.0 documentation*. [online] pillow.readthedocs.io. Available at: <https://pillow.readthedocs.io/> [Accessed 19 Oct. 2025].

Prichard, E.C. (2021). Is the Use of Personality Based Psychometrics by Cambridge Analytical Psychological Science's 'Nuclear Bomb' Moment?. *Frontiers in Psychology*, [online] 12. doi: <https://doi.org/10.3389/fpsyg.2021.581448> [Accessed 14 Sep. 2025].

PyPi (2019). *opencv-python*. [online] PyPI. Available at: <https://pypi.org/project/opencv-python/> [Accessed 18 Oct. 2025].

Python Software Foundation (2024). *sqlite3 — DB-API 2.0 interface for SQLite databases — Python 3.8.2 documentation*. [online] docs.python.org. Available at: <https://docs.python.org/3/library/sqlite3.html> [Accessed 19 Oct. 2025].

Python.org (n.d.). *difflib — Helpers for computing deltas*. [online] Python documentation. Available at: <https://docs.python.org/3/library/difflib.html#difflib.SequenceMatcher> [Accessed 19 Oct. 2025].

Renotte, N. (2021). *How to Install OpenCV for Python // OpenCV for Beginners*. [online] YouTube. Available at: <https://www.youtube.com/watch?v=M6jukmppMqU> [Accessed 18 Oct. 2025].

Requests.readthedocs.io (n.d.). *Requests: HTTP for Humans™ — Requests 2.31.0 documentation*. [online] requests.readthedocs.io. Available at: <https://requests.readthedocs.io/> [Accessed 19 Oct. 2025].

SBS News (2018). *'Likely' Australians caught up in Cambridge Analytica data scandal*. [online] SBS News. Available at: <https://www.sbs.com.au/news/article/likely-australians-caught-up-in-cambridge-analytica-data-scandal/bnh3h7olz> [Accessed 14 Oct. 2025].

Schulz, J. (2013). Geometric optics and strategies for subsea imaging. [online] doi: <https://doi.org/10.1533/9780857093523.3.243> [Accessed 19 Oct. 2025].

Shah, M. (2024). *Fanning the Flames: Online Misinformation and Far-Right Violence in the UK - GNET*. [online] GNET. Available at: <https://gnet-research.org/2024/08/28/fanning-the-flames-online-misinformation-and-far-right-violence-in-the-uk/> [Accessed 9 Oct. 2025].

Socratica (2023). *SQLite in Python || Python Tutorial || Learn Python Programming*. [online] www.youtube.com. Available at: <https://www.youtube.com/watch?v=c8yHTlrs9EA> [Accessed 18 Oct. 2025].

SQLite (2014). *Datatypes In SQLite Version 3*. [online] Sqlite.org. Available at: <https://www.sqlite.org/datatype3.html> [Accessed 19 Oct. 2025].

Sutton, J. (2025). *Big Five Personality Traits: The OCEAN Model Explained [2019 Upd.] *. [online] PositivePsychology.com. Available at: <https://positivepsychology.com/big-five-personality-theory/> [Accessed 2 Oct. 2025].

Tesseract-ocr.github (2020). *Tesseract User Manual*. [online] tessdoc. Available at: <https://tesseract-ocr.github.io/tessdoc/> [Accessed 19 Oct. 2025].

The New York Times (2017). Presidential Election Results: Donald J. Trump Wins. *The New York Times*. [online] 9 Aug. Available at: <https://www.nytimes.com/elections/2016/results/president> [Accessed 3 Sep. 2025].

The News API (2025). *Free live and top story JSON news API | The News API*. [online] Thenewsapi.com. Available at: <https://www.thenewsapi.com/> [Accessed 19 Oct. 2025].

Urbansky, D. (2022). *World News API. * [online] Worldnewsapi.com. Available at: <https://worldnewsapi.com/> [Accessed 19 Oct. 2025].

Woods, L. (2024). *Disinformation and disorder: the limits of the Online Safety Act. * [online] Online Safety Act Network. Available at: <https://www.onlinesafetyact.net/analysis/disinformation-and-disorder-the-limits-of-the-online-safety-act/> [Accessed 9 Oct. 2025].

GitHub: <https://github.com/abmiah/msc-dissertation-code>

Video Demo: <https://www.youtube.com/watch?v=aCb6T-xMF7w>