Luisa Stracqualursi
Patrizia Agati

DIGITAL COMMUNITIES AND COLLECTIVE BEHAVIOUR

Data, Personality, and Polarisation in the Age of Artificial Intelligence

Bologna
University Press

Biblioteca

Luisa Stracqualursi, Patrizia Agati

DIGITAL COMMUNITIES AND COLLECTIVE BEHAVIOUR
Data, Personality, and Polarisation in the Age of Artificial Intelligence

Bologna
University Press

Il volume è stato pubblicato con il contributo dell'Alma Mater Studiorum - Università di Bologna e del Dipartimento di Scienze Statistiche "Paolo Fortunati" dello stesso Ateneo.

Fondazione Bologna University Press
Via Saragozza 10, 40123 Bologna
tel. (+39) 051 232 882

www.buponline.com
e-mail: info@buponline.com

Il testo pubblicato è stato sottoposto a peer review

ISBN 979-12-5477-715-2
ISBN online 979-12-5477-716-9
DOI 10.30682/9791254777169

Questo volume è stato realizzato a partire da un impaginato camera-ready in formato pdf fornito dall'autore

In copertina: immagine generata tramite SORA AI

Prima edizione: dicembre 2025

TABLE OF CONTENTS

List of abbreviations

ABSA	Aspect-Based Sentiment Analysis
AI	Artificial Intelligence
API	Application Programming Interface
Bi-GRU	Bi-directional Gated Recurrent Unit
Bi-LSTM	Bi-directional Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
DT	Decision Tree
DL	Deep Learning
GB	Gradient Boost
GMM	Gaussian Mixture Model
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
KDE	Kernel Density Estimation
KNN	K-Nearest Neighbour
LDA	Latent Dirichlet Allocation
Light-GBM	Light Gradient Boosting Machine
LIWC	Linguistic Inquiry and Word Count (dictionary)
LLaMa	Large Language Model Meta AI
LLM	Large Language Model
LR	Logistic Regression
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
MBTI	Myers Briggs Type Indicators
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
NB	Naïve Bayes
NLP	Natural Language Processing
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
PLSA	Probabilistic Latent Semantic Analysis
RF	Random Forest
RoBERTa	Robustly optimized BERT pretraining approach
ROC	Receiver Operating Characteristic

SA	Sentiment Analysis
SAD	Statistical Analysis of Distributions
SNA	Social Network Analysis
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
VADER	Valence Aware Dictionary and sEntiment Reasoner
XGB	Extreme Gradient Boosting

Introduction

Digital communities have become a privileged observatory for the quantitative study of human behaviour. The continuous production of textual, visual, and audio content, together with the intricate web of user interactions, generates traces that enable large-scale analyses of opinions, relational dynamics, and information processes.

This volume arises from the need to provide a coherent statistical framework for observing, measuring, and interpreting such phenomena with tools suited to their nature: massive, multimodal (textual and visual), dynamic, and above all intrinsically noisy.

The starting premise is twofold. On the one hand, the data generated by digital communities represent an unprecedented informational resource in terms of breadth and granularity; on the other, their abundance does not in itself guarantee validity or representativeness. Data access, platform and content biases, and the choice of models employed—every stage of the information pipeline, from collection to modelling—introduces decisions with concrete inferential consequences.

The approach of this monograph is that of a methodological *scoping review*: it offers a critical survey of the main techniques as reported in the peer-reviewed literature, highlighting trade-offs, assumptions, and metrics, with the aim of providing an operational roadmap to make the best use of digital traces.

More specifically, the text is organised into three parts conceived as a unified design. The first part addresses the processing and collection of digital data: from the types of data available to the challenges of

access in what is often described as the "post-API era"[1]. In recent years, restrictive policies of several platforms have limited access through their official APIs, amplifying the resort to web scraping, which in turn raises technical and legal issues. This part also examines the principal techniques of data cleaning and preprocessing, with a statistical focus on how apparently "technical" operational choices can significantly affect the outcomes of analyses, and how such effects may be mitigated through robust procedures.

The second part focuses on polarisation. The argument here is clear: polarisation is not imposed by a pervasive algorithm but emerges from the interaction between user exposure choices, homophily (the tendency to connect with similar others), and recommendation systems which, by optimising engagement, modulate the ordering and visibility of content. The analysis unfolds across three dimensions: struc*ture* (through 'community detection' techniques and modularity Q), *content* (via topic modelling and sentiment analysis, from lexical approaches to contextual models), and the *shape of distributions* (identifying J-shaped patterns, estimating densities/mixtures, and applying bimodality tests such as Hartigan's *dip test*). The chapter also includes a critical discussion of the limitations and conditions of use of these metrics.

The third part connects personality traits with observed online behaviours. Starting from the lexical hypothesis and major psychometric models (Big Five and MBTI), the text retraces the evolution of estimation approaches: from early lexical methods to classical machine learning models, and up to contextual embeddings and the most recent Transformer-based architectures, discussing predictive effectiveness and trait-specific challenges. Personality, estimated through language, is then linked to collective dynamics. What emerges is a picture in which micro and macro meet: individual profiles help explain aggregate patterns of polarisation, while collective structures in turn shape individual expression.

Regarding the use of large language models (LLMs), the stance adopted is measured. Their exploratory and simulation capabilities are

[1] API (Application Programming Interface): an interface that enables structured and automated access to the data and functionalities of a digital platform, according to the rules and limitations established by the service provider.

acknowledged—for instance, in generating scenarios or testing hypotheses—yet it is emphasised that all results must remain anchored to observed data, statistical controls, and careful assessment of risks of hallucination or amplification of pre-existing biases. LLM-assisted analysis is considered a valuable support for investigation, provided it is embedded in a methodological framework grounded in empirical evidence and statistical validation.

The intended readership includes statisticians, data scientists, and social scientists interested in working with digital data.

On the ethical plane, the work adopts a clear position: public availability of data does not equate to unrestricted freedom of use. Responsible research on digital communities requires protocols that take into account users' reasonable expectation of privacy.

Ultimately, what is proposed here is a bridge: between data collection and inference, between the individual and the collective, between computational power and methodological discipline. A bridge that makes it possible to traverse the dense and imperfect world of digital communities without getting lost in the noise, to distinguish signal from illusion, and to transform vast quantities of data into reliable knowledge.

Part I

DATA COLLECTION FROM DIGITAL COMMUNITIES

Digital communities — social groups that interact primarily through online platforms — have become central elements of contemporary life. These communities are deliberately formed around shared interests, goals, or identities, and are characterised by the continuous and spontaneous production of textual, visual, and audio content, as well as by interactions that generate extensive sets of digital traces. These user-generated digital data, produced spontaneously through online activity, represent a valuable resource for statistically analysing opinions, behaviours, and emerging social dynamics on a large scale.

The statistical study of data generated by digital communities offers unprecedented opportunities for understanding social phenomena in real time. Using appropriate analytical techniques — such as Natural Language Processing (NLP) for textual content, Computer Vision for visual data, and advanced neural models for audio — it is possible to detect emerging trends, monitor shifts in public opinion, and identify complex relational and social patterns that would be difficult to access using traditional data collection methods.

However, the collection and use of such data present significant challenges that must be carefully addressed. First, there are issues related to data availability and accessibility, primarily shaped by the restrictive policies imposed by digital platforms. While platforms may provide access tools such as APIs, these are increasingly limited, making it difficult to collect complete and representative data in a systematic way. This situation — sometimes referred to in the literature as the 'post-API era' — has led many researchers to adopt alternative methods such

as web scraping, which in turn raise further technical difficulties and legal ambiguities.

In addition to access-related challenges, data collection from digital communities introduces specific methodological problems, including various forms of bias. These include *sampling bias*, due to the non-representativeness of accessible data; *platform bias*, stemming from the socio-demographic characteristics of users on specific platforms; *inclusion bias*, the systematic over- or under-representation of particular categories of users or content; and *content bias*, resulting from the selective and partial nature of self-disclosed experiences and opinions online (see Tab.1). If not properly addressed through rigorous statistical techniques, these biases may undermine both internal validity (i.e., the correct identification of relationships among variables) and external validity (i.e., the generalisability of results to other populations or contexts).

Table 1. Main sources of bias in digital community data collection

Type of bias	Definition	Methodological implications
Sampling bias	Arises from the non-representativeness of accessible data.	Limits inferential validity; may distort observed distributions and relationships.
Platform bias	Derives from the socio-demographic characteristics of users on a given platform.	Reduces comparability across platforms; may confound behavioural patterns with user composition.
Inclusion bias	Systematic over- or under-representation of particular categories of users or content.	Skews the weight of perspectives, interactions, or topics; may amplify marginal or dominant groups.
Content bias	Results from the selective and partial nature of self-disclosed experiences and opinions.	Produces incomplete or distorted signals; threatens both internal and external validity.

Furthermore, the use of digital data from online communities raises important ethical concerns regarding privacy, informed consent, and the effective anonymisation of participants. Although such data may

be publicly accessible, this does not necessarily imply that users have consciously consented to their use for research purposes, especially in the case of large-scale aggregation. It is therefore essential to adopt strict ethical standards and implement data processing procedures that ensure user privacy and adhere to the core principles of ethical research.

This chapter systematically addresses each of these issues. It begins by defining the concept of digital community, distinguishing it from earlier notions of virtual communities, and proposes a detailed taxonomy of the main types of digital data generated within such groups. It then outlines the primary quantitative techniques used by researchers to collect these data — namely, API access and web scraping — analysing their strengths, technical limitations, and legal implications. The chapter proceeds by discussing the strategic and epistemic value of digital data for the social and applied sciences, highlighting the specific advantages they offer over traditional data collection methods. Subsequently, it examines the main biases that threaten the quality of statistical results and offers concrete methodological strategies to address and mitigate them. Finally, it provides an in-depth discussion of the most relevant ethical concerns, drawing on established guidelines for conducting responsible and user-respectful statistical research.

In sum, this chapter aims to provide a clear, precise, and rigorous overview of the processes involved in collecting and using data from digital communities, highlighting both their substantial statistical and analytical potential and the need for methodologically and ethically conscious approaches. Only by doing so will it be possible to fully and appropriately leverage the invaluable informational wealth embodied by contemporary digital communities.

1. Digital Communities and User-Generated Data

1.1 From Virtual to Digital Communities

Since the early days of the Internet, scholars and pioneers have sought to define the concept of **virtual community**, laying the

theoretical groundwork for understanding online social interactions. Rheingold (1993) described these communities as "*social aggregations that emerge from the Net when enough people carry on public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace.*" This initial definition — based on the experience of WELL, one of the first online communities — emphasised the spontaneous formation of authentic bonds in a cyberspace perceived as separate from the physical world.

Manuel Castells (2000, 2002) later expanded on this perspective, defining virtual communities as "*self-defined electronic networks of interactive communication organised around a shared interest or purpose*". This definition introduced elements of intentionality and structure, underlining how online communities are conscious networks based on common goals and sustained by communicative interaction.

As the Internet spread, terminology also evolved. The term "virtual community" was criticised for being too generic and for implying that such communities were somehow less real or tangible. Kindsmüller et al. (2010), for instance, preferred the term **online community**, which clearly indicates the medium (online vs. face-to-face), without questioning the actual reality of the community. From this perspective, an online community is always real, as its members are aware of their belonging. More recently, the term **digital community** has become widespread. Although initially perceived as more technical — referring directly to technological and computational features — it is now widely used to refer to communities characterised predominantly by interactions mediated by digital devices, often in combination with offline ones.

Over time, the boundary between online and offline worlds has become increasingly blurred. In the 1990s, virtual communities were often perceived as parallel and separate realities compared to everyday life. Today, digital communities represent a natural extension of everyday social relationships, transcending what has been defined as "digital dualism," that is, the clear-cut separation between online and offline life. For many contemporary users, who have grown up with social media and permanent Internet access, the distinction between "real" and "virtual" interactions has lost meaning: friendships, debates, and

feelings of belonging flow seamlessly through both digital and non-digital channels.

Digital communities have now become central arenas for significant social phenomena such as collaborative learning, the formation of professional networks, and social activism. This terminological evolution directly reflects the sociocultural shift that has taken place. In the 1990s, the term virtual community referred to the Internet as a separate space, accessed by a relatively limited group of users. With the rise of Web 2.0[2] and the widespread diffusion of permanent connectivity, the concept of online community highlighted the genuine nature of mediated interactions. Today, the term digital community conveys the ubiquity and normalisation of digital technology, which has become an integral and inseparable part of the contemporary social fabric.

Understanding this terminological and conceptual evolution is essential to correctly contextualise the statistical analysis of data generated by contemporary digital communities.

1.2 Types of Data in Digital Environments

Digital communities generate a vast and diverse flow of data that researchers can exploit to statistically analyse preferences, opinions, social networks, and online cultural dynamics. To effectively conduct such analyses, it is essential to understand the main categories of data available online. These data can be grouped into four main macro-categories, each of which can be further subdivided into specific types:

- *Content data*: This includes everything that users actively create and share (posts, comments, images, videos, audio, etc.). The analysis of such data requires specific techniques such as Natural Language Processing (NLP) for text, deep learning for images, and speech recognition for audio, all of which convert the data into numerical formats suitable for statistical analysis.

[2] Web 2.0 refers to the evolution of the Internet in the early 2000s, characterised by the proliferation of interactive platforms that allow users to generate, share, and comment on content online in a participatory manner (e.g., social networks, blogs, wikis, video platforms).

- *Interaction data*: These capture how users engage with both content and one another (views, likes, friendships, shares, etc.). Analysing these data is essential to study community structure and information virality, often through techniques such as Social Network Analysis (SNA).
- *Metadata*: These are contextual information that enrich content and interaction data. They include temporal data (timestamps) and spatial data (geolocation). Metadata are crucial for conducting longitudinal analyses or identifying geographical differences in online social dynamics.
- *User data*: This category refers to information about user profiles and self-declared identities, such as age, gender, interests, professions, account creation date, etc. Although useful for understanding participant characteristics, these data must be interpreted with caution due to the risk of inaccuracy and low statistical representativeness.

The detailed taxonomy proposed in the following section is intended to offer researchers an organised and systematic framework for effectively collecting, organising, and analysing data from digital communities. Understanding the nature and specificities of each data type is in fact essential for setting up sound methodological studies. For example, a study on hate speech might focus primarily on textual content and user interaction dynamics, while research on disinformation networks might instead target content redistribution chains, temporal metadata (message synchronisation), and information related to suspicious accounts.

As Weller et al. (2013) point out, adopting a well-structured taxonomy not only enables more complete and organised data collection, but also facilitates the appropriate selection of data acquisition techniques (e.g., selecting APIs dedicated to content or metadata) and supports early consideration of ethical issues. For instance, analysing sensitive user profile and demographic information requires significantly more caution than the analysis of public data such as tweets or open posts.

1.2.1 Content Data

These data include all materials that users actively create and share, forming the core content that circulates within digital communities. Based on format, they can be divided into:

- *Textual data*: Every form of written communication produced by users (posts, tweets, status updates, comments, captions, hashtags, etc.) is considered textual data. As unstructured data, they cannot be analysed directly with traditional statistical methods and require transformation into numerical format through NLP techniques. A common procedure is *vectorisation*, which converts text into numerical vectors. Methods such as *Bag of Words* or *TF-IDF* generate vector representations in which each element corresponds to the frequency or relative importance of a word. These vectors can then be used in statistical classification algorithms or in topic modelling techniques to identify the main themes within a textual dataset.
- *Visual data*: These consist of images, photographs, memes, GIFs, and videos. Statistical analysis of such data relies on their conversion into numerical variables through deep learning algorithms, such as Convolutional Neural Networks (CNNs). These networks process image pixels and automatically extract relevant features (shapes, colours, objects, facial expressions), which can then be used as variables in statistical analyses. Common applications include counting the frequency of specific objects or symbols in image collections, analysing colour schemes or facial expressions to study cultural phenomena, or monitoring the evolving visual representation of events.
- *Audio data*: These include voice messages, podcasts, and sound clips. Statistical analysis of audio data uses techniques such as speech recognition and prosodic analysis (rhythm and intonation) to extract acoustic features. Such features—for

instance, *Mel-Frequency Cepstral Coefficients* (MFCCs)[3]—can be treated as numerical variables and employed in statistical models and advanced neural networks, such as CNNs, *Long Short-Term Memory* (LSTM) networks, or *Transformers* (Vaswani et al., 2017), such as *Whisper* or *wav2vec* [4]. Analysing audio al-lows researchers to explore patterns in spoken language, emo-tional tone, and the emergence of musical trends or *sound memes* within communities.

1.2.2 Interaction Data

These data capture how users interact both with content and with each other, representing the social actions that connect individuals and shape information diffusion. Three main subtypes can be identified:

- *Engagement with content*: This includes metrics and actions that indicate users' direct interaction with a specific piece of content. Typical examples are the number of views or "likes", which can be treated as quantitative variables for statistically analysing a content's popularity. Other actions, such as emoji reactions or upvotes, serve as indicators of appreciation or immediate emotional response. Comments, in contrast, represent more articulated textual feedback and can be analysed using NLP techniques.
- *User-to-user engagement*: This refers to actions reflecting direct social ties and interactions between community members. This category includes "follow" or friendship relations, which indicate persistent social connections and personal networks.

[3] Mel-Frequency Cepstral Coefficients (MFCCs) are numerical parameters that describe the acoustic properties of speech, particularly timbre. They are widely used to analyse and compare vocal signals in a compact and biologically plausible manner.

[4] A Transformer is a specific type of neural network introduced by Vaswani et al. (2017). Models such as *Whisper* and *wav2vec* are neural networks based on the Transformer architecture, designed for automatic speech recognition. They combine the effectiveness of deep learning with attention mechanisms to analyse and transcribe audio signals in a highly accurate and flexible manner.

For instance, analysing follow relationships on Twitter or Instagram enables mapping the social network and identifying influential users or cohesive groups. It also encompasses mentions and tags—that is, explicitly referencing another user in a conversation—which signal mutual interaction and visibility, as well as participation in specific groups or communities. Theoretically, direct messages (DMs) also belong to this category, but they are generally inaccessible to researchers for ethical and privacy reasons. User-to-user interaction data can be analysed primarily through Social Network Analysis (SNA) techniques, which allow researchers to study community structure, identify subgroups, key roles (such as influencers or brokers connecting otherwise disconnected clusters), and relational dynamics (e.g., mutual support, conflict).

- *Content redistribution*: This includes actions through which users disseminate content created by others (e.g., shares, retweets, reposts). Statistical analysis of these data is fundamental for tracing content virality and studying information cascades. Reconstructing sharing networks enables researchers to analyse, for example, how a piece of news (or a fake news item) becomes viral, which users act as amplification hubs due to their large followings, and how the propagation of information or disinformation unfolds across the social network.

1.2.3 Metadata

Metadata are contextual information related to the types of data mentioned earlier (content and user-linkage data). They can be categorised as follows:

- *Temporal data*: These consist of *timestamps*, indicating the exact moment of each interaction. Essential for longitudinal analyses, these data allow the study of temporal dynamics, the identification of activity spikes, and the detection of seasonal trends. Timestamps also serve as the foundation for building time series and forecasting models.

- *Spatial data*: These include geolocation information associated with posts or user profiles. They can be used in spatial analyses to map discussions and behaviours, detect regional differences, or examine the geographic diffusion of specific phenomena. Typically, they are treated as categorical variables (e.g., city, region) or as geographical coordinates in geostatistical models.
- *Technical data*: These refer to information about the devices used by users (e.g., mobile or desktop), operating systems, and browsers. Although seemingly less directly connected to content, they are highly useful for identifying usage patterns. For example, a platform might show mobile usage spikes at certain times of the day and desktop usage at others, suggesting different usage contexts (leisure vs. work). Additionally, they help identify anomalous activity: an account sending hundreds of messages per day via automated APIs from a Linux server may indicate a bot, in contrast to human users typically accessing via smartphones with more natural rhythms. Technical data thus contribute to improving dataset quality by supporting the identification and removal of spambots[5] or duplicate accounts and offer valuable insights into digital divide–related phenomena[6].

1.2.4 User Data

This category concerns information related to users' profiles and self-declared identities within digital communities. Such information allows researchers to link individual user characteristics to their online activity, providing a more comprehensive framework for interpreting social dynamics in digital environments. Three main subtypes can be identified:

---

[5] A spambot (a contraction of spam and robot) is an automated program designed to send or publish unsolicited messages in bulk on digital platforms.

[6] The digital divide refers to the inequality in access to, skills in, and use of digital technologies among individuals, groups, or geographical areas.

- *Demographic data*: These include basic information such as age, gender, language, and sometimes self-reported ethnicity or country of origin. When available, demographic data support statistical comparisons between user groups, enabling researchers to examine, for instance, how online habits vary by age range or how certain content resonates differently across genders. However, these data should be interpreted with caution: when not directly verified by the platform, they may not accurately represent real users. Moreover, online samples often do not reflect the general population (e.g., older adults tend to be underrepresented on many platforms).

- *Profile information*: This refers to what users choose to publicly share in their bios or "About" sections, such as interests, professions, and affiliations. These data offer useful context for identifying interest-based communities or for interpreting users' content and interactions from a statistical perspective.

- *Account details*: These are technical and behavioural attributes associated with user accounts, such as account creation date, number of followers, and activity frequency (e.g., average number of posts per day). Such indicators allow for distinguishing between different user types—for example, new versus long-time accounts, or influencers versus *lurkers* (i.e., users who consume content without actively engaging). When combined with interaction data, account details also help identify anomalous behaviour. For instance, a recently created account with a high follower count may suggest suspicious activity such as follower purchasing.

It is important to note that demographic and profile data are often self-declared and rarely verified by an external authority, which may result in inaccuracies or deliberate falsifications. Users may provide incorrect information intentionally—to protect their privacy, as a joke, or to present a distorted image of themselves (e.g., claiming to be experts in a topic to gain credibility). For this reason, the statistical representativeness of demographic and profile data cannot be taken for granted. For example, a study comparing online behaviours across age

groups may yield results that are not representative of reality if based solely on unverified, self-reported data. Therefore, any statistical study relying on such unverified information should interpret results cautiously and explicitly acknowledge the methodological and statistical limitations in its conclusions.

2. Quantitative Approaches to Data Extraction and Preprocessing

The quantitative approach aims to collect large volumes of structured data (*big data*) in order to identify general trends, correlations, and robust statistical patterns. Two main techniques are used by researchers to obtain such data: the use of official platform-provided *Application Programming Interfaces* (APIs) and *web scraping*. There are also other complementary quantitative methods, such as online surveys. However, these are generally characterised by smaller sample sizes (typically hundreds or a few thousand respondents) compared to those obtained through APIs or web scraping, and they also require longer data collection times, as participants must actively complete the surveys. For these reasons, the present section focuses primarily on automatically extracted data obtained via APIs or scraping directly from digital platforms.

2.1 Data Collection via APIs

APIs, offered directly by social platforms (such as X [formerly Twitter], Facebook, Instagram, Reddit), allow researchers to acquire data in a structured and authorised manner (Wilson, 2022). Originally designed for commercial purposes, APIs enable developers to interact with platform databases under predefined conditions. Although API use is methodologically sound—providing data that are already structured and ready for statistical analysis—their availability is subject to restrictions imposed by the platform providers. These restrictions often concern the maximum volume of accessible data, the types of data available, and the frequency of allowed requests.

A notable example is the *Academic Research API* introduced by Twitter in 2021, initially praised for its generous access to historical data, but later significantly restricted in 2023 following the platform's acquisition and rebranding as "X" (Bastos, 2025). Currently, X's academic API offers only a minimal volume of data compared to the past, making it difficult to conduct robust analyses that require access to large-scale historical or real-time data.

This instability in API-based access may introduce systematic biases into the scientific literature, favouring studies that focus on more open platforms like Reddit, to the detriment of closed or highly moderated environments, which are less accessible (Burnat & Davidson, 2025). At the same time, inequalities emerge within the scientific community itself: large institutions with significant financial resources or privileged partnerships with platforms (e.g., 'Social Science One' with 'Meta/Facebook') enjoy preferential access compared to independent researchers or smaller research centres (Tromble, 2021). Such methodological limitations can negatively affect the statistical representativeness of collected samples (Bosch et al., 2025a).

In response to these issues, new regulatory frameworks such as the *Digital Services Act* (DSA, EU Regulation 2022/2065) are attempting to impose greater transparency obligations on platforms regarding data sharing with academic researchers, although their practical implementation remains to be clearly defined.

From an operational perspective, using APIs typically involves the following steps:

- *Access request*: Researchers must usually register as developers with the platform of interest and obtain an API key, subject to approval by the platform itself.
- *Query formulation*: Requests must follow the official technical documentation provided by each platform. For instance, X's API (formerly Twitter) requires the use of specific *endpoints* [7](e.g., `/tweets/search/recent` for recent tweets,

[7] An "endpoint" is typically a specific URL to which HTTP requests are sent in order to retrieve data or perform defined operations. Each endpoint is

/tweets/counts/recent for time-based counts) and defined parameters (e.g., keywords, hashtags, specific users, time intervals).

- *Data retrieval*: The data obtained via APIs are returned in structured, standardised formats, primarily JSON (*JavaScript Object Notation*), which can be easily converted into datasets for statistical analysis using programming languages such as Python or R.

For example, Reddit's official API is commonly used through *PRAW* (Python Reddit API Wrapper), a Python library that facilitates structured and automated access to data from Reddit communities. Compared to other platforms, it offers relatively generous limits, allowing up to around one thousand items per request (data verified as of May 2025). It enables researchers to systematically access content published in public subreddits, including detailed information on posts (author, content, number of votes, comments) and their temporal dynamics. Similarly, the *Graph API* by Facebook (now Meta) allows—subject to specific authorisation—the retrieval of information on pages, public posts, and engagement metrics such as comments, reactions, and shares, which are useful for studies on user behaviour or viral information diffusion.

In general, while technically effective, API use requires careful methodological attention and a sound understanding of platform-specific limitations, in order to ensure data quality, representativeness, and the replicability of statistical analyses.

2.2 Data Collection via Web Scraping

When access to data through APIs is unavailable or insufficient, many researchers resort to *web scraping*. This technique involves the use of automated scripts to extract content directly from web pages, simulating the behaviour of a human user (Luscombe et al., 2022).

---

associated with clearly defined functions and is documented in the official API documentation, which also specifies the required parameters and the response format.

Compared to APIs, web scraping offers greater versatility and independence from the corporate decisions of platform owners. However, it also presents numerous technical challenges. Frequent changes in website architecture, the implementation of anti-scraping systems such as CAPTCHAs or mandatory logins, and dynamic interfaces based on JavaScript can significantly disrupt or complicate data collection (Bastos, 2025).

Advanced software tools such as *Scrapy* (available only in Python) and *Selenium* (usable in Python, R, and other languages) make it possible to address these complexities, but they also introduce further challenges. These include the risk of overloading website servers, the need for technical pauses between requests, and the use of IP proxies to avoid permanent blocking.

A crucial limitation of web scraping lies in the *difficulty of retrieving complete historical data*. Typically, only the data currently visible on a website can be extracted, leaving potentially significant temporal gaps that may compromise the methodological validity of longitudinal analyses (Mimizuka et al., 2025). Moreover, web scraping is necessarily *limited to public content*, generating a significant sampling bias since private or closed-group data are not accessible (Bosch et al., 2025a).

While APIs—being generally compliant with platforms' terms of service—offer greater legal and ethical security, web scraping involves *more legal ambiguity*. Although some legal precedents, such as the "*hiQ Labs vs. LinkedIn"* case in the USA (2019)[8], have upheld the right to collect publicly available data via scraping, many platforms still consider this activity a violation of their terms of service. At the ethical level, the debate centres on users' "reasonable expectation of privacy", particularly with regard to content that is technically public but perceived as semi-private (Trezza, 2023).

In summary, the quantitative approach—whether through APIs or web scraping—offers powerful tools for the collection and analysis of large-scale data. However, the effectiveness of these tools depends heavily on the availability and quality of the data itself. The instability

[8] (*hiQ Labs, Inc. Vs. LinkedIn Corp., Sentence No. 17-16783*, 2019)

of API access and the technical and ethical challenges of web scraping require researchers to maintain a critical awareness in order to ensure scientific rigour, privacy protection, and methodological accuracy in their conclusions.

3. Data Cleaning and Preprocessing

The statistical analysis of data from digital communities, particularly textual content, requires a careful phase of cleaning and preprocessing. Texts collected online are intrinsically noisy: they may contain typographical errors, acronyms, emojis, links, mentions, hashtags, non-standard punctuation, and heterogeneous formatting. The goal of cleaning and preprocessing is to transform raw text into a coherent representation free from redundancies, while preserving all relevant information. Operational choices regarding which techniques to adopt vary according to the downstream algorithm or model: in this work a distinction is made between a **classical pipeline**, designed for sparse representations[9], and a **minimal pipeline**, optimised for contextual models[10] or lexicon-sensitive sentiment models[11]. The selection of one or the other directly influences the cleaning and normalisation phases.

3.1 Handling Missing Data and Outliers

In big data (e.g., millions of tweets), records with missing or empty text are usually deleted without imputation. Outliers in textual variables (e.g., exceptionally short or long texts) are generally not handled on a

[9] Sparse representations: vector-based encodings of text (e.g., bag-of-words, TF-IDF) characterised by high dimensionality and a large proportion of zero entries, where each dimension corresponds to the presence or frequency of specific terms.

[10] Contextual models: neural language models (e.g., BERT, DistilBERT) that generate word representations dynamically based on surrounding context, capturing semantic and syntactic nuances beyond fixed vocabularies.

[11] Lexicon-sensitive sentiment models: sentiment analysis approaches that rely on predefined lexical resources, where the presence and intensity of specific words (e.g., positive or negative terms) directly influence sentiment scores.

case-by-case basis, since in very large datasets their overall statistical impact tends to be negligible. Instead, only basic technical filters are applied, such as removing duplicates or discarding degenerate strings with fewer than 2–3 meaningful words, always documenting the criteria adopted. In general, depending on the type of analysis, it is advisable to set a minimum (and in some cases maximum) threshold of useful words — for instance, excluding tweets with fewer than 10 words or more than 500 for topic modelling — below or above which the text is considered respectively informationally insufficient or redundant.

For ancillary **numerical variables** (e.g., number of shares or likes), if the proportion of missing values is negligible, deletion is appropriate; otherwise, robust imputation techniques are preferable, such as replacement with the median, possibly computed by group or period. Extreme values in numerical variables may be detected using the *Interquartile Range* (IQR) and either replaced or treated directly with *winsorisation*[12]. Recall that the IQR is the difference between the third quartile and the first quartile. A numerical value x is considered an outlier if $x < Q_1 - k \cdot IQR$ or $x > Q_3 + k \cdot IQR$, where $k = 1.5$ for "moderate" outliers and $k = 3$ for "extreme" outliers.

Example: if $Q_1 = 10$, $Q_3 = 40$, then $IQR = 30$. With $k = 1.5$, the thresholds are -35 and 85; with $k = 3$, the thresholds are -80 and 130. Values beyond these cut-offs are classified as outliers.

Once thresholds are calculated via IQR, it is possible to "exclude" values beyond them (*trimming*), thereby reducing the dataset but eliminating genuine outliers (useful when extremes reflect clear data collection errors), or to replace extreme values with the calculated thresholds (this replacement constitutes winsorisation, with thresholds derived from quartiles)[13].

---

[12] Winsorisation: a technique named after the biostatistician Charles P. Winsor (Winsor, 1932), consisting in the replacement of extreme values with the nearest observations within a specified percentile range.

[13] For highly skewed count variables (e.g., number of likes, shares), winsorisation may be replaced or complemented by logarithmic transformations $\log(1 + x)$ or by robust regression methods (Huber & Ronchetti, 1981; Koenker & Bassett Jr, 1978), which reduce the impact of extreme values without introducing sharp cut-offs.

Winsorisation consists in capping extreme values: observations below a lower threshold (e.g., 1st percentile) are replaced with the threshold itself, and similarly for those above an upper threshold (e.g., 99th percentile). Unlike trimming, observations are not deleted; rather, extremes are limited while preserving sample size and relative order within the distribution. Example: if the 1st percentile of likes is 0 and the 99th percentile is 5,000, all values <0 become 0, and all values >5,000 become 5,000. Typically, IQR-based rules are applied when exclusion criteria are required, whereas winsorisation is preferable when one wishes to attenuate the impact of outliers without reducing sample size (robust means/regressions). Trimming is advisable only when extremes clearly reflect artefacts rather than real phenomena.

3.2 Text Cleaning and Normalisation

Text cleaning involves non-linguistic operations aimed at removing or transforming elements that introduce noise or are irrelevant for analysis. Common operations include:

a) Treatment of non-lexical elements. URLs, mentions, and hashtags can be removed or replaced with neutral tokens (e.g., <URL>, <USER> instead of leaving the full link or username). These neutral tokens have no direct semantic value, but preserve useful structural information: the model recognises the presence of a link or mention without being influenced by its specific text. Hashtags can be transformed into words by removing the hash symbol or decomposed into subwords (e.g., «#DataScience» → «Data Science»). Residual HTML markup should also be removed.

b) Punctuation, emojis, and numbers. Punctuation and emojis often carry relevant affective signals, such as exclamation marks, emoticons, or elongated letters and punctuation marks. In sentiment tasks — whether with lexicon-based models such as VADER (Valence Aware Dictionary and sEntiment Reasoner, Hutto & Gilbert, 2014) or contextual language models such as BERT or DistilBERT (Sanh et al., 2020) — it is advisable to retain them, as they contribute to meaning and emotional intensity. In sparse representations, such as bag-of-words or TF-IDF, ordinary punctuation is usually removed to reduce

noise and dimensionality. Numbers in the text may be retained, normalised into a neutral token (e.g., «<NUM>»), or removed, depending on the analytical objectives.

c) Removal of personal data. It is recommended to mask any email addresses, phone numbers, or IP addresses present in the text with neutral tokens, for ethical and compliance reasons, avoiding the storage of personal data in plain form.

Normalisation, in contrast, is the process of standardising linguistic forms, aimed at reducing lexical variance and ensuring that equivalent variants of the same word are treated identically. Again, choices should be guided by the analytical objective, calibrated to the type of analysis or model to be applied; otherwise, useful signals may be removed or noise left in place, potentially biasing estimates. Typical operations include:

1. *Lowercasing and redundant spacing***.** Lowercasing facilitates uniformity in classical lexical models; for VADER and Transformer-based models, which exploit orthographic and casing signals, it is preferable to avoid it. Normalising spacing improves the quality of subsequent tokenisation.
2. *Tokenisation.* This is the division of text into elementary units (tokens), generally words or subwords, according to language-sensitive rules.
3. *Stopwords.* Removal of very frequent, low-information words (articles, prepositions, conjunctions). This reduces dimensionality and noise in sparse representations (e.g., TF-IDF, bag-of-words), but should be avoided when contextual meaning is important, as in contextual models (BERT/DistilBERT).
4. *Lemmatisation and stemming.* Lemmatisation reduces a word to its canonical form (lemma) based on morphological rules and dictionaries (e.g., «walking», «walked», «walks» → «walk»). Stemming truncates a word to its root using simple rules, without ensuring a correct word (e.g., all forms reduced to «walk» or even «wal»). Stemming is fast but approximate, whereas lemmatisation yields canonical forms more useful for

semantic comparison and topic modelling. For Transformer-based models, such techniques are generally superfluous and sometimes counterproductive.

It is therefore advisable to distinguish between a *classical pipeline* (lowercasing, punctuation removal, stopword removal, lemmatisation) for sparse representations, and a *minimal pipeline* (preserving casing, punctuation, emojis, stopwords, and non-destructive normalisation of links and mentions) for contextual and lexicon-sensitive sentiment models (see Tab. 2).

Table 2. Comparison of classical and minimal preprocessing pipelines

Preprocessing step	**Classical pipeline (sparse representations)**	**Minimal pipeline (contextual/lexicon-sensitive models)**
Lowercasing	Applied	Not applied (casing preserved)
Punctuation	Removed	Preserved
Emojis	Removed	Preserved
Stopwords	Removed	Preserved
URLs, mentions	Removed or replaced with neutral tokens	Replaced with neutral tokens (non-destructive)
Hashtags	Removed or transformed into words	Transformed into words (subword decomposition if needed)
Numbers	Removed or normalised	Preserved or normalised as neutral tokens
Lemmatisation	Applied	Not applied
Stemming	Sometimes applied (fast but approximate)	Not applied

To ensure traceability and reproducibility, it is necessary to document: the chosen pipeline (classical or minimal) and the adopted options for URLs, mentions, hashtags, punctuation, and emojis; criteria for handling missing data; thresholds for outliers (IQR/quantiles, minimum word counts); whether lemmatisation or stemming was used;

and versions of language packages and models. The description of preprocessing choices must accompany results, as they directly affect the internal and external validity of subsequent analyses.

4. Strategic Value of Data from Digital Communities

Despite the difficulties associated with accessing and collecting online data, the data generated by digital communities retain strategic value for the social and applied sciences. Several studies have emphasised these advantages over traditional approaches such as surveys or other more static data collection techniques (González-Bailón, 2017; Lazer et al., 2021; Salganik, 2019).

First and foremost, online platforms generate continuous streams of real-time data, allowing for *immediate and dynamic monitoring* of emerging social phenomena (Gayo-Avello, 2013). This makes it possible to promptly capture the evolution of users' opinions and behaviours, significantly reducing common issues of traditional methods, such as recall bias in survey responses (Jungherr et al., 2020).

Another key advantage lies in the *cost-effectiveness* and broad geographic reach of digital analyses. Compared to traditional methods such as face-to-face interviews or telephone surveys, digital data allow researchers to reach large and diverse populations rapidly and at relatively low cost (Salganik, 2019). Social groups that are typically hard to engage in research—such as patients with rare diseases or stigmatised minorities—can be included through existing digital communities, thereby improving the representativeness and inclusiveness of the analysed samples (Lupton, 2020).

Finally, the digital context facilitates the *development of longitudinal studies*, as it allows for continuous data collection over time. This enables the systematic monitoring of behavioural or attitudinal changes, enhancing the temporal and statistical validity of research findings (González-Bailón, 2017).

These strategic advantages translate into a wide range of practical applications—from marketing and social listening (Humphreys & Wang, 2018), to the monitoring of health-related misinformation, the

analysis of collaborative learning on online educational platforms (Wise & Paulus, 2016), and the design of public policies based on data drawn directly from the communities involved.

Despite the strategic value of data derived from digital communities, several critical issues may compromise their scientific validity. Various authors have warned about the risk of *big data hubris*, that is, the mistaken belief that a large quantity of data automatically equates to validity and quality (Jungherr et al., 2020). A prominent example of the risks associated with overconfidence in big data is Google Flu Trends, which, by relying exclusively on Google search queries, overestimated influenza cases for several weeks. This case illustrates how big data hubris—treating big data as a substitute rather than a complement to traditional methods—can undermine the validity of analysis. (Lazer et al., 2014).

Among the most frequently observed biases is *sampling bias*, which refers to distortions arising from samples that do not adequately represent the target population (Horvát & Hargittai, 2021; Tufekci, 2014). This may occur through *self-selection* by users who voluntarily choose to produce publicly accessible content, or through *undercoverage*, when certain demographic groups or user categories primarily interact in environments that are inaccessible to researchers, such as private groups, closed communities, or restricted areas of platforms.

A second relevant issue is *platform bias*, caused by the demographic and behavioural characteristics specific to each social platform (Blank & Lutz, 2017). For example, data collected from Twitter—a platform that tends to overrepresent younger, urban populations—may lead to misleading conclusions if generalised to broader populations. Similarly, Reddit tends to host a user base with higher-than-average technological competence, thus generating systematic biases if used as the sole data source for generalisable studies. Furthermore, user behaviour differs significantly between anonymous platforms (Reddit, 4chan) and identity-based platforms (Facebook, LinkedIn), producing additional behavioural biases (Barbosa & Milan, 2019).

A third methodological concern is *content bias*, i.e., the tendency of users to provide selective or partial representations of their opinions or experiences online (Boyd & Crawford, 2012). This behaviour leads

to distorted portrayals of reality, often emphasising either positive or negative aspects, and limits the generalisability of findings (Marwick & Boyd, 2014).

These biases directly impact both the *internal* and *external validity* of studies. Internal validity is compromised when observed correlations between variables are misinterpreted as causal relationships, neglecting the possible influence of confounding factors[14].External validity is limited when the study's conclusions cannot be generalised to other contexts or populations due to the specific characteristics of the sample or platform analysed (Salganik, 2019).

Finally, the *reliability* (replicability) of data collected via APIs or scraping may be undermined by sudden changes in platform policies (Freelon, 2018; Bruns, 2021), which make it difficult to ensure stable conditions for data collection over time.

To effectively address these challenges, appropriate corrective strategies must be adopted, including statistical weighting techniques to compensate for sample distortions, the integration and comparison of data from multiple platforms (*data triangulation*), and a critical analysis of the socio-demographic characteristics of the user base (Lazer et al., 2021).

Only by adopting such a rigorous approach can researchers fully harness the epistemic potential of digital community data, ensuring the scientific robustness and credibility of their conclusions.

[14] A confounding factor (or confounder) is a "hidden" variable that can account for an apparent relationship between two phenomena, leading to a misinterpretation of results. For example, consider a study that finds people who drink more coffee are more likely to have heart problems. However, a potential confounder may be that those who consume more coffee also tend to smoke more. In this case, the true cause of the increased heart problems would be smoking (the confounding factor), not coffee itself. In short, failing to account for confounders can lead to attributing a causal relationship to a variable when no such relationship actually exists.

5. The Ethical Dimension

Statistical research based on digital community data inevitably raises fundamental ethical questions, primarily concerning the definition of public versus private spheres, informed consent, and the protection of anonymity. Statistical analyses must adhere to established principles of research ethics, such as those outlined in the Belmont Report (Department of Health, Education, and Welfare., 1979), which call for respect for persons, beneficence (i.e., minimisation of harm), and justice in participant selection (Markham & Buchanan, 2012).

One of the key challenges is clearly defining the boundary between public and private in digital environments. The collection of publicly accessible data (e.g., tweets or open comments on online forums) is generally treated as public observation and, in most cases, does not require individual informed consent (Franzke et al., 2020). However, even when working with public data, researchers must consider users' *reasonable expectation of privacy*, particularly when data contain sensitive information that could personally identify individuals.

Informed consent, a cornerstone of ethical research, becomes problematic in large-scale statistical analysis, where obtaining explicit individual consent is logistically unfeasible. In such cases, it is advisable to use data in aggregated and anonymised form, ensuring that no individual can be identified directly or indirectly through variable combinations (Bosch et al., 2025b).

Anonymisation, however, presents specific risks in digital research. The use of pseudonyms or simple removal of identifiers is not always sufficient to ensure privacy, as information can often be re-identified through search engines or cross-referencing with external datasets. Therefore, advanced *de-identification* techniques are recommended, such as *generalisation* and *controlled perturbation* of information, and avoiding the inclusion of searchable verbatim quotes (Franzke et al., 2020; Harris et al., 2024).

To support consistent and rigorous ethical decisions, the literature recommends relying on established institutional frameworks, such as those of the *Association of Internet Researchers* (AoIR) and the *National Committee for Research Ethics in the Social Sciences and the Humanities*

(NESH, 2021). Both provide operational guidelines based on a situational risk assessment approach, incorporating parameters such as information sensitivity, participant vulnerability, and the level of interaction between researcher and subject.

In summary, ethics in statistical research on digital communities requires the rigorous application of privacy safeguards, effective anonymisation, and responsible use of public data, in line with well-established approaches recommended by the international scientific community.

Part II

POLARISATION IN DIGITAL COMMUNITIES

This chapter addresses the complex phenomenon of polarisation within digital communities, beginning with a fundamental question: "*Are digital platforms merely mirrors reflecting existing societal divisions, or are they powerful engines that accelerate and reshape them*?" It argues that the answer lies in a complex interplay between human psychology and technological architecture. The central thesis is that polarisation is not imposed by an omnipotent algorithm but rather co-created. This process is driven by a feedback loop: our natural tendency toward homophily[15] leads us to engage with similar content; algorithms, in turn, detect this behaviour to maximize engagement, subsequently providing us with similar material and reinforcing our isolation within a "cocoon" that we ourselves contribute to weaving.

To empirically analyze and measure this phenomenon, the following sections present a rigorous methodological framework based on several complementary quantitative pillars:

– *Social Network Analysis* (SNA): Examines the topology of interactions to identify the formation of distinct communities. Techniques such as community detection based on modularity maximization (Q) allow quantification of the strength of these divisions.

[15] Homophily: the tendency for individuals to associate and bond with others who are similar to themselves ("like attracts like").

- *Sentiment Analysis* (SA): Extends beyond traditional lexical methods (such as VADER) towards advanced contextual and deep learning models like BERT, capable of capturing affective polarisation with greater precision and contextual sensitivity in textual content.
- *Statistical Analysis of Distributions* (SAD): Utilizes formal statistical tests, such as Hartigan's "dip test," to detect bimodality (an indicator of polarisation) in opinion data. Additionally, the chapter analyzes emblematic cases of "J-shaped" distributions in online reviews using density estimation techniques and mixture models to quantify divergences in judgments.
- *Topic modelling*: Employs advanced Natural Language Processing (NLP) techniques, including Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), and transformer-based methods such as BERTopic and Top2Vec, to uncover and categorize the major themes within digital discourse. This helps elucidate the ideological narratives underpinning polarized groups.
- *Artificial Intelligence (AI) and Large Language Models (LLMs)*: Highlights recent advancements in AI, specifically focusing on transformer-based Large Language Models such as GPT-4, Gemini, Claude, and Llama. These recent models facilitate additional analyses beyond traditional methods, such as 'ideological positioning', which quantitatively assesses how textual data aligns with specific political ideologies, and the simulation of social network dynamics through synthetic networks and autonomous agents, enabling an in-depth examination of the mechanisms underlying polarisation.

1. The Fragmentation of Consensus in the Digital Era

The contemporary digital ecosystem increasingly represents an arena of fragmented debates, marked by polarized and often irreconcilable positions. To deeply understand this dynamic, it is essential to rigorously clarify certain key concepts. Specifically, we distinguish two closely related but distinct phenomena:

- *Ideological or Group Polarisation*, where deliberation among like-minded individuals leads to collective positions more extreme than those the individuals would have taken independently (Sunstein, 2002).
- *Affective Polarisation*, which describes an intensification of antipathy and distrust towards opposing groups. This phenomenon goes beyond mere disagreement on political issues, encompassing increased animosity, distrust, and hostility towards any opposing group (Iyengar et al., 2019). In recent years, particularly in countries like the United States, affective polarisation has grown faster than ideological polarisation (Druckman et al., 2021).

To explore how these forms of polarisation are amplified in digital environments, we analyze the psychological and technological mechanisms that foster and sustain this process, focusing specifically on interactions between individual cognitive predispositions and platform algorithmic structures.

At the heart of this phenomenon is a fundamental sociological principle: *homophily*, the tendency of "like to attract like." The inclination to form connections with individuals sharing our beliefs, values, and backgrounds is a foundational organizing principle of human societies (McPherson et al., 2001). Digital platforms, however, amplify this principle on an unprecedented scale and with unmatched efficiency. The ease with which we can find and connect with niche groups of similar individuals worldwide facilitates the formation of homogeneous communities with previously unimaginable density and reach.

Building upon this psychological foundation are platform architectures, giving rise to two frequently discussed concepts: "*echo chambers*" and "*filter bubbles*". Echo chambers primarily represent a social phenomenon rooted in group polarisation (Sunstein, 2001, 2002; L. Kim, 2023). Within these chambers, beliefs are amplified and reinforced through continuous repetition within closed systems, isolating individuals from counter-narratives and opposing evidence. Filter bubbles, conversely, specifically refer to the intellectual isolation resulting from algorithmic personalization, which selects the content users see,

creating a unique and potentially limited informational environment for each user (Pariser, 2011). For example, recommendation systems and automatic content suggestions frequently steer users toward content similar to what they have previously engaged with, further reinforcing isolation.

However, scientific research compels us to move beyond a simplistic understanding of these phenomena. A foundational study conducted on Facebook platform revealed a surprising result: users' individual choices (who they follow, what they click on) constitute a more substantial factor limiting exposure to diverse content than the platform's ranking algorithms themselves (Bakshy et al., 2015a). In other words, users themselves are the primary curators of their own bubbles. This result has been further supported by analyses conducted across multiple platforms, confirming that user homophily is the main driver of online segregation, although the specific architecture of each platform can modulate the intensity of this effect (Cinelli et al., 2021).

The popular narrative of the "filter bubble" tends to depict the user as a passive victim of an omnipotent algorithm. Scientific evidence, instead, offers a far more complex picture — one of a co-created feedback loop. Polarisation is not something algorithms impose upon us; rather, it is something we *co-create* with them. The process unfolds as follows: driven by a natural tendency toward homophily, users choose to interact with content that confirms their views. The algorithm, designed to maximize engagement, observes this behaviour and correctly deduces that users "prefer" such content. Consequently, it provides more similar material — not due to a malicious intent to divide but because this strategy is most effective at keeping users on the platform. Presented with a feed full of similar content, users feel validated, interact further, and become increasingly isolated, thereby reinforcing the signals sent to the algorithm. The result is not a top-down imposed "bubble," but rather a "cocoon" actively woven by users themselves, efficiently amplified by technology.

2. Traditional Statistical Methodologies for the Analysis of Polarisation

To detect and measure polarisation within digital communities, researchers employ a wide array of quantitative methods. Given that no single technique can entirely capture the complexity of the phenomenon, a holistic approach combining structural network analysis and textual content analysis is indispensable.

2.1 Social Network Analysis (SNA) and Communities

Social Network Analysis (SNA) is a fundamental methodology for studying polarisation by examining the structure of interactions (e.g., replies, retweets, mentions) among users. Structurally, a polarised network is characterised by the formation of distinct, densely interconnected communities with sparse connections between them (Interian et al., 2023). SNA allows researchers to transcend the analysis of individual attitudes, instead visualising the collective structure of a debate by identifying influential nodes, information cascades, and the overall topology of division.

More specifically, digital polarisation can be measured by analysing the social networks connecting actors (users, accounts) based on their interactions. Formally, a network is modelled as a graph $G = (V, E)$ with $|V| = n$ nodes (users) and $|E| = m$ edges (relationships between users) (see Fig. 1).

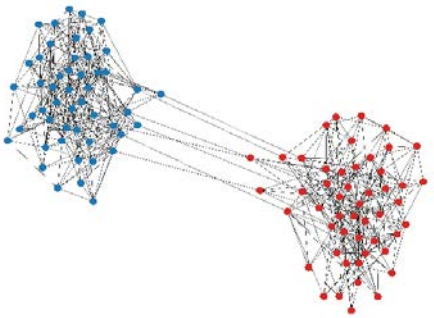

Figure 1- Polarised network

Various centrality metrics allow quantifying the structural importance of nodes within the graph (Freeman, 1977, 1978). The primary metrics are presented below:

- **The degree centrality** of a node measures how directly a node is connected with other nodes in the network. Formally, it is defined as:

$$C^{\deg}(i) = \frac{k_i}{n-1}, \tag{1}$$

where k_i is the degree of the node i, that is, the number of direct links with other nodes, and n is the total number of nodes. $C^{\deg}(i)$ is obtained precisely by comparing the node's degree to the maximum value it can assume, which is $(n-1)$, corresponding to the case in which the node is directly connected to all other nodes in the network. A node with high degree centrality (also known as a *hub*) can exert strong direct influence within the network.

- The **closeness centrality** indicates how quickly a node can reach all other nodes in the network. The formula is:

$$C^{close}(i) = \frac{n-1}{\sum_{j \neq i} d(i,j)}, \tag{2}$$

where $d(i,j)$ is the minimum geodesic distance (the shortest path length) between nodes i and j. High values indicate nodes capable of rapidly disseminating information throughout the entire network.

- The **betweenness centrality**, measures how frequently a node acts as a "bridge" on shortest paths between pairs of other nodes. For undirected graphs, it can be written as:

$$C^{betw}(i) = \frac{\sum_{k \neq j \neq i} \frac{\sigma_{kj}(i)}{\sigma_{kj}}}{(n-1)(n-2)/2}, \tag{3}$$

where:

- σ_{kj} is the total number of shortest paths between nodes k and j;

- $\sigma_{kj}(i)$ represents how many shortest paths pass through node i.

- The summation considers all distinct pairs of nodes j and k, both different from node i.

This metric thus defined takes values between 0 and 1 depending on whether the node i is located on all or none of the shortest paths between pairs of nodes. For directed graphs, the denominator of the formula (3) becomes $(n-1)(n-2)$, as each pair of nodes is distinctly considered in both possible directions. In a polarised network, a node with high betweenness can connect separate communities, serving as an information bridge (broker) in the network. In polarised contexts, betweenness centrality is particularly relevant, as it identifies nodes linking otherwise separated communities (e.g., users bridging two factions)

An essential technique in SNA is **community detection**, which is used to identify and analyse subgroups (or "communities") within a social network. The most widely adopted approach is based on the optimisation of a function known as *modularity* (Q), which quantifies the strength of the division of a network into communities (Girvan & Newman, 2002; Newman & Girvan, 2004). Formally, given a partition of nodes into communities $C_1, C_2, ..., C_h$, modularity measures the difference between the fraction of edges that fall within these communities and the fraction expected if the edges were distributed randomly within the network.

For undirected and unweighted graphs, modularity is defined as:

$$Q = \frac{1}{2m} \sum_{i,j} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \tag{4}$$

where:

- a_{ij} is a generic element of the adjacency matrix $\mathbf{A}$, taking the value 1 if an edge exists between nodes i and j, and 0 otherwise.
- k_i and k_j are the respective degrees of the nodes i and j.
- m represents the total number of edges in the network (the cardinality of the edge set).

- $\delta(c_i, c_j)$ equals 1 if nodes i and j belong to the same community and 0 otherwise (Newman, 2006).

A positive value of Q close to 1 indicates a strong community structure (i.e., significantly more intra-community edges than expected by chance), whereas values close to 0 or negative values suggest the absence of meaningful communities. In polarised networks, one expects high modularity, indicative of cohesive communities (e.g., ideological groups) separated by few reciprocal connections.

In practice, modularity values exceeding 0.3 indicate a robust community structure, where nodes are significantly more likely to be connected within their own community than to nodes belonging to other communities (Kopacheva & Yantseva, 2022). Maximising modularity is thus a powerful method to identify echo chambers and polarised factions within a network. Common algorithms for modularity maximisation include the *Fast Greedy* algorithm (implemented, for example, in the Python library "igraph") and *spectral optimisation methods* (available in the "Scikit-learn" library). These algorithms identify the network partition that yields the highest modularity Q, thus revealing the most robust and statistically significant community divisions (Alzahrani et al., 2021; Jiang & Xu, 2023).

Specific algorithms exist to identify the community structure that maximizes modularity Q. Among the most prominent and widely applied is the **Girvan–Newman algorithm** (Girvan & Newman, 2002), which operates by iteratively removing edges that serve as connections between different communities. Specifically, the algorithm computes a measure called "edge betweenness", which corresponds to the number of shortest paths passing through a given edge. Edges with high betweenness values typically link densely connected subnetworks; thus, removing them maximizes the separation among communities. By iteratively recalculating edge betweenness after each removal, the network progressively splits into increasingly smaller and tighter communities, forming a hierarchical community structure that can be represented through a dendrogram, ranging from the most general (macro) to the most detailed (micro) levels (Fig. 2).

The Girvan–Newman method yields a hierarchical sequence of partitions, from which the optimal division is obtained by selecting the

level with maximum modularity Q. However, the computational cost is high, as repeatedly recalculating betweenness values makes the method computationally demanding for large networks.

For large-scale analyses, more efficient algorithms such as the **Louvain method** (Blondel et al., 2008) are preferred. The Louvain method adopts a *greedy* approach, optimizing modularity locally. Initially, each node constitutes a separate community; subsequently, through successive iterations, nodes are reassigned to the neighbouring cluster that produces the highest increment in modularity Q until no further local improvements can be achieved. The identified communities are then aggregated into "super-nodes," and the process is repeated on a reduced graph until convergence is reached. This multi-level algorithm exploits the fact that small local optimizations of modularity rapidly lead to a good global partition, making it highly scalable even for networks containing millions of nodes.

In summary, the Louvain method maximises the modularity index Q in an approximate but computationally efficient manner, whereas the Girvan–Newman method provides an exact hierarchical solution but with greater computational demands. Both approaches are employed to detect communities within social networks: Girvan–Newman is particularly useful for exploring multi-level structures, while Louvain efficiently produces high-modularity partitions on large-scale data.

From an implementation perspective in Python, the "NetworkX" library provides functions for computing all standard centrality metrics (*degree, closeness, betweenness*), and it includes an implementation of the Girvan–Newman algorithm, based on iterative edge removal. For the Louvain method, external libraries (e.g., "python-louvain") or graph analytics tools such as "iGraph" can be employed. These packages offer functions for rapid community clustering, returning both the resulting partitions and the corresponding modularity value Q.

These SNA techniques allow researchers to quantify topological *polarisation* by examining, for instance, whether the interaction network divides into coherent modules (echo chambers) characterised by sparse reciprocal connections—a scenario typical of highly polarised communities.

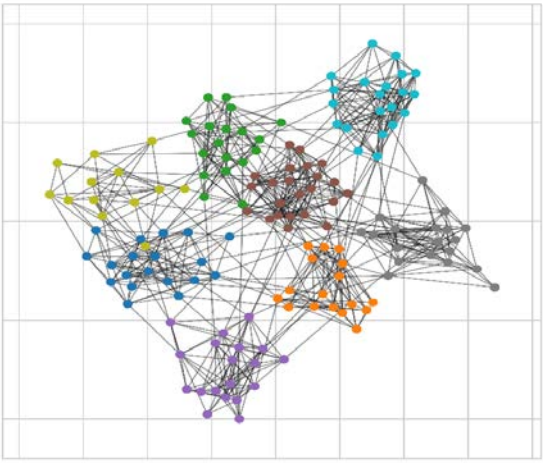

Figure 2 – Network with modularity Q=0.748

However, it is important to be aware of a well-known methodological limitation inherent in approaches based on modularity maximisation: the so-called *resolution limit* (Fortunato & Barthélemy, 2007). This intrinsic property of the modularity metric Q can hinder algorithms—particularly in large-scale networks—from identifying actual communities smaller than a certain scale. In practice, the algorithm might merge multiple small, clearly defined clusters into a larger community if this merging results in a greater increase in the global modularity value. In the context of polarisation studies, this limitation implies a risk of failing to detect niche "echo chambers", potentially grouping them into broader ideological clusters, and thus losing a high-resolution view of network fragmentation.

2.2 Statistical Analysis of Distributions

When analysing quantitative opinion data (e.g., aggregated sentiment scores, ideological indices, ratings, etc.), a common indicator of polarisation is the presence of a bimodal or multimodal distribution (two or more distinct peaks), rather than a *unimodal* distribution (a

single peak). To rigorously assess whether a distribution significantly deviates from unimodality, one can employ Hartigan's test for unimodality, also known as the **dip test** (Hartigan & Hartigan, 1985).

This test provides a quantitative measure of the "dip," defined as the maximum vertical distance between the empirical cumulative distribution of the data and the closest unimodal cumulative distribution. Formally, given a sample of n observations with empirical distribution function $F_n(x)$, the dip statistic D_n is defined as follows:

$$D_n = \inf_{F \in U} \sup_x \left| F_n(x) - F(x) \right| \tag{5}$$

where U denotes the set of all possible unimodal cumulative distributions. In other words, D_n measures the smallest maximum vertical distance achievable between the empirical distribution $F_n(x)$ and any theoretical unimodal cumulative distribution $F(x)$.

Therefore, large values of D_n indicate that the empirical distribution significantly deviates from any unimodal distribution, suggesting the presence of multiple clusters within the data (potential polarisation into groups). Hartigan and Hartigan also provided reference tables and analytical approximations to assess the statistical significance of the dip test, using as the null hypothesis the uniform distribution on the interval, which represents a limiting case of a "flat" unimodal distribution.

To assess distributional shape, Figure 3 provides an example of unimodal and bimodal distributions, highlighting how the dip test can be employed to evaluate the null hypothesis of unimodality.

In practice, the test returns a p-value associated with D_n: a very small *p*-value (e.g., $p<0.01$) leads to the rejection of the null hypothesis that the data come from a unimodal distribution, thus supporting the presence of bi- or multimodality. In the context of polarisation, this indicates a high probability that distinct subgroups exist within the data (e.g., clearly separated "pro" vs "against" opinions).

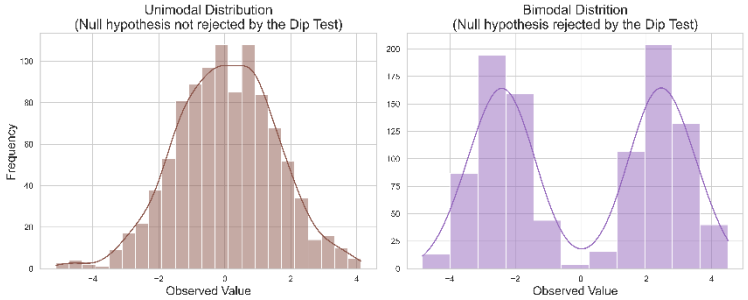

Figure 3 - Unimodal vs Bimodal Distributions (Hartigan's Dip Test)

From a computational perspective, the *dip test* requires an efficient algorithm to find the "closest unimodal distribution" to the empirical data. The authors proposed a *grid search* procedure on the ordered data, optimized with isotonic regression techniques[16], which allows for the calculation of the dip statistic, D_n, in $O(n^2)$ time or better.

Ready-to-use software implementations are available today. In R, the *diptest package* is a popular choice, while in Python, implementations like the *diptest module* on PyPI (which integrates C++ code for efficiency) or the *unidip* library can be used. For example, *unidip* can quickly compute the dip value and even recursively identify multimodal intervals (peak detection) in the data.

It is worth noting that the dip test is a non-parametric and distribution-free test; it does not assume a specific shape for the modes. This is a key advantage over methods like the Silverman test, which requires kernel density estimation (Silverman, 1981).

This makes it suitable for exploratory analysis of polarisation. By applying it to the distribution of sentiment within two political groups, or to ratings on a 1-5 scale, one can objectively verify whether the aggregate distribution is compatible with a single center (representing consensus or moderation) or if it instead presents multiple distinct poles (a sign of polarisation).

---

[16] Isotonic regression is a non-parametric optimization technique used to enforce a monotonicity constraint between variables. This means the dependent variable is required to increase (or decrease) monotonically with respect to the independent variable (Barlow, 1972).

In combination with visualizations (such as histograms or estimated density plots) and other metrics, the dip test provides a rigorous criterion for declaring the presence of *statistical polarisation* in a distribution of opinion data.

2.3 J-Shaped Distributions in Online Reviews: Estimation and Measures of Polarisation

A paradigmatic case study of distributional polarisation is found in online reviews expressed as ratings (e.g., a 1–5 star scale). It has been observed that such evaluations often do not follow a normal, centrally-located curve; instead, they exhibit a strongly skewed *J-shaped distribution* (see Fig. 4).

In a typical J-shaped distribution of reviews, the majority of ratings accumulate at the positive extreme (5 stars), while a smaller portion gathers at the negative extreme (1 star), with a scarcity of intermediate ratings (2–4 stars). Large-scale studies (e.g., analyses of millions of Amazon reviews) quantitatively confirm this phenomenon: over 70% of ratings tend to be ≥4 stars, with a secondary residual peak of minimal ratings and an almost empty "hole" in the middle (N. Hu et al., 2007). This distribution is *asymmetric* and *bimodal* (one main peak and a secondary one), thus violating the assumption of symmetric unimodality that would justify the naive use of the arithmetic mean as a summary statistic. Consequently, the mean of ratings under a J-shaped distribution becomes a biased indicator of perceived product quality. This occurs because the mean is inflated by self-selection bias: typically, users who are either highly enthusiastic or strongly dissatisfied are more likely to leave reviews, while those who are moderately satisfied tend to remain silent. Therefore, quantifying and statistically modeling a J-shaped distribution is crucial for correcting these biases and extracting more reliable measures of polarisation from online reviews (Roh & Yang, 2021).

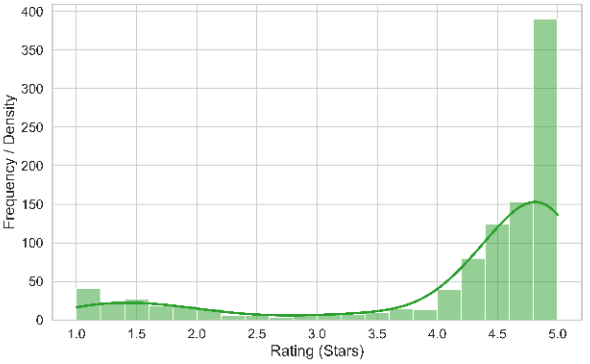

Figure 4 - J-shaped distribution of online reviews on a 1–5-star scale

A useful initial tool is *Kernel Density Estimation* (KDE), which enables continuous visualisation and analysis of the shape of a rating distribution (Hastie et al., 2001; Parzen, 1962; Rosenblatt, 1956; Silverman, 1981). Given a series of scores $x_1,...,x_n$, KDE produces a smoothed density function $\hat{f}(x)$ by placing a symmetric "kernel" (e.g., Gaussian) of bandwidth h around each data point. Formally:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{6}$$

where K is a unimodal kernel function, usually Gaussian, and h is the bandwidth. By adjusting h, KDE provides a smoothed estimate of the underlying distribution: a small bandwidth captures fine details (potentially overfitting noise), whereas a large bandwidth yields a smoother curve that may obscure genuine multimodality. In the context of online reviews, an appropriate KDE typically reveals a high density peak near the maximum score (e.g., 5 stars) and a smaller one near the minimum (1 star), thereby confirming a J-shaped distribution.

KDE is implemented in various Python packages: for instance, "seaborn" provides the *kdeplot* function for visualisation; "sklearn.neighbors.KernelDensity" enables numerical computation of $\hat{f}(x)$; and "statsmodels" includes advanced tools for bandwidth selection (e.g., Silverman's rule, Scott's rule, cross-validation).

These tools help quantify polarisation by indicating whether the density function exhibits multiple distinct modes. For example, one

can estimate the probability $P(X \in \text{"}extremes\text{"})$ by integrating $\hat{f}$ on x near the minimum and maximum values, and compare it to $P(X \in \text{"}centre\text{"})$: in a J-shaped distribution, the probability mass in the extremes is typically much higher than in the centre, reflecting strong opinion polarisation.

Moreover, *local maxima* of the estimated density (i.e., points where the first derivative changes from positive to negative) can be used to objectively identify separate clusters of ratings—typically corresponding to "lovers" vs "haters" of a product.

Beyond non-parametric estimation, a model-based approach involves representing a bimodal distribution as a mixture of two or more subdistributions. Statistically, this is achieved using *finite mixture models*, e.g., assuming ratings are generated by two Gaussian components: one centred at a high value (e.g., $\mu \approx 4.5$) and one at a low value ($\mu \approx 1.5$). A two-component Gaussian mixture model (GMM) has the form:

$$f(x) = \pi \cdot \mathrm{N}\left(x \mid \mu_1, \sigma_1^2\right) + (1-\pi) \cdot \mathrm{N}\left(x \mid \mu_2, \sigma_2^2\right) \tag{7}$$

where π denotes the proportion of extremely positive reviewers (Mignemi et al., 2024). Fitting such a model using the Expectation-Maximisation (EM) algorithm allows estimation of the parameters $\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ that best explain the observed data. In the case of a truly J-shaped distribution, the EM algorithm tends to converge toward well-separated means μ_1 and μ_2, that are distant from the overall mean, and on a mixing proportion, π, that is proportional to the observed skewness. (e.g., $\pi \approx 0.7$, heavily weighted toward the dominant positive component).

One advantage of mixture models is that they provide a *direct quantification of polarisation*: for example, a polarisation index can be defined as the normalised distance $|\mu_1 - \mu_2|$, or based on the degree of overlap between the components (less overlap implies stronger polarisation). In computational social science, such mixture-based approaches (including nonparametric variants) are commonly used to identify latent clusters of divergent raters (Martinková et al., 2023). For instance, Bayesian mixture models (e.g., based on a Dirichlet process) can identify lenient vs strict *raters* in educational assessments, introducing a

latent polarisation index λ based on the separation between posterior distributions (Koudenburg & Kashima, 2022).

In the case of online reviews, a simple GMM implemented with "sklearn" is often sufficient. One can estimate the optimal number of components k using model selection criteria such as AIC or BIC—typically, $k = 2$ is appropriate—and assign each reviewer a probability of belonging to either cluster (e.g., enthusiast vs critic). This enables more refined analyses, such as testing whether specific user types (e.g., "Vine" participants[17] or "top reviewers" [18]) are more likely to belong to the dominant positive component.

Mixture models thus provide a probabilistic framework for quantifying polarisation, complementary to the descriptive insights offered by KDE: while KDE visually highlights the presence of multiple peaks, mixture models estimate their separation and relative weights.

A J-shaped distribution is also characterised by *strong right-skewness*. In addition to detecting multimodality, it is often useful to formally test whether the distribution deviates from symmetry (e.g., relative to the median). Several statistical tests are available for this purpose. One of them is the test by Miao et al., which tests for symmetry around an unknown median using robust statistics (Miao et al., 2006) (median and Median Absolute Deviation) rather than classical measures (mean and standard deviation), improving upon the Hotelling–Solomons approach (Hotelling & Solomons, 1932).

In general, many symmetry tests rely on the null condition $F(m+\Delta) = 1 - F(m-\Delta),\ \forall \Delta$, i.e., mirror symmetry around a median m, and build test statistics around this property (e.g., the Cabilio–Masaro test, the symmetric Kolmogorov–Smirnov test) (Cabilio & Masaro, 1996; Milošević & Obradović, 2019).

In Python, there are no built-in symmetry test functions, but a practical approach is to compute the *empirical skewness* and test its

[17] Vine Voices are a select group of reviewers invited by Amazon to receive free products from sellers in exchange for honest and unbiased reviews. Their reviews are marked with a special badge.

[18] Top Reviewers were a group of influential reviewers who were ranked by Amazon based on the helpfulness votes their reviews received. This program has since been discontinued, and the ranking system is no longer active.

significance. The "scipy.stats.skewtest" function tests the null hypothesis $H_0 : skewness = 0$.(symmetry) against $H_1 : skewness \neq 0$. When applied to online ratings, this test typically yields a highly significant result (e.g., p-value $\ll 0.05$), indicating that the distribution is strongly asymmetric.

Another informative metric is *kurtosis*, which measures the "tailedness" of the distribution. Polarised or bimodal distributions may exhibit either high or low kurtosis, depending on how the probability mass is distributed in the tails. In the J-shaped case, kurtosis tends to be high, due to pronounced tails (many ratings of 1 and 5 relative to fewer intermediate values).

Modelling the *J-shaped distribution* of online reviews therefore requires a combination of statistical tools: KDE for visualising and estimating the empirical shape, mixture models for quantifying latent components and their separation, statistical tests (e.g., for unimodality and symmetry) to validate the significance of the observed polarisation, and summary indices to compare different datasets (e.g., different products or time periods) based on their degree of polarisation. When integrated into a coherent framework, these tools allow researchers to rigorously characterise digital polarisation in online reviews—not only by showing that the distribution is J-shaped, but also by quantifying the **intensity** of the polarisation (e.g., the size of the gap between extremes, the relative weight of opposing factions, or the departure from normality).

Python offers a wide range of tools for implementing these analyses: from "numpy" and "scipy" for basic statistics, to "matplotlib" and "seaborn" for visualisation, to "Scikit-learn" for mixture modelling. These allow for fully replicable and scalable studies on large review datasets. The combined use of these statistical methods ensures a robust and scientifically grounded approach to analysing digital polarisation, with results that are both interpretable and theoretically aligned with established literature.

2.4 Additional Quantitative Measures of Distributional Polarisation

Beyond tests and models, several summary metrics have been proposed to quantify *how much* a distribution is bimodal or polarised. A classical measure is **Sarle's bimodality coefficient**, denoted as b

(Sarle, 1983). This index combines the sample skewness (γ_3) and kurtosis (γ_4) and is defined as:

$$b = \frac{\gamma_3^2 + 1}{\gamma_4 + \frac{3(n-1)^2}{(n-2)(n-3)}} \tag{8}$$

For a perfectly symmetric unimodal distribution (e.g., Gaussian), b typically falls below 0.55, whereas values exceeding $5/9 \approx 0.555$ suggest potential bimodality. The coefficient reaches its maximum value of 1 in the limiting case of a Bernoulli distribution concentrated on two values only (i.e., maximum polarisation). Conversely, distributions with heavy tails but a single central peak may yield low values of b, despite exhibiting high variance.

In the context of online reviews, computing b from the data provides an interpretable scalar indicator: for instance, if a product's ratings yield $b = 0.62$, this serves as strong evidence of bimodality (i.e., polarised evaluations). In Python, the bimodality coefficient can be calculated by obtaining skewness and kurtosis (e.g., via "scipy.stats.skew" and "scipy.stats.kurtosis"; the latter returns *excess kurtosis*, so 3 must be added to obtain γ_4).

An additional family of polarisation measures, this time from political economics, is the **Esteban-Ray polarisation index**, often cited in opinion studies (Esteban & Ray, 1994). In essence, the index attains its maximum when the population is evenly split into two equally sized and well-separated groups, and it decreases as individuals either spread across more positions or remain within a single cluster. Although the original formula for continuous populations is mathematically complex, it conceptually combines a measure of inequality with a measure of group identification. In the discrete case of two groups, the index increases with both the distance between group means and the balance in group sizes (with a 50/50 split maximising polarisation). While not natively implemented in Python, this index can be computed by following the specifications provided by the authors, and has been employed, for example, to measure political polarisation in voting distributions and ideological opinions.

Finally, a simpler yet effective indicator is the **extreme polarisation ratio** R, defined as:

$$R = \frac{\#\text{rating(max)} + \#\text{rating(min)}}{2} \tag{9}$$

This expresses the proportion of *extreme ratings* relative to the *total*, capturing the idea that in a polarised distribution, most observations fall at the extremes. Values close to 1 indicate high polarisation, while low values suggest moderation (prevalence of mid-scale ratings). Naturally, this rudimentary metric cannot distinguish between a truly bipolar (bimodal) distribution and one where extreme ratings are uniformly spread, but when used in combination with other tools (e.g., KDE, dip test), it serves as a helpful summary statistic (DiMaggio et al., 1996).

2.5 Sentiment Analysis (SA)

To understand the sentiments and underlying ideas of digital polarisation, one of the most widely used techniques today is *Sentiment Analysis* (SA), which identifies the orientation and emotional content expressed in textual data.

Established tools for sentiment analysis using lexical approaches include models such as VADER (*Valence Aware Dictionary and sEntiment Reasoner*), specialized in analysing social media texts (Hutto & Gilbert, 2014), and *SentiStrength* (Thelwall et al., 2010), designed specifically for short and informal texts such as SMS, online comments, and brief social media posts. In literary contexts or, more broadly, for longer narrative texts, the *Syuzhet* package (Jockers, 2015) provides a more articulated and versatile approach, allowing the use of multiple emotion lexicons including the NRC Emotion Lexicon (Mohammad & Turney, 2010), the AFINN lexicon (Nielsen, 2011), and the Bing lexicon (M. Hu & Liu, 2004).

For all lexicon-based models, *accurate preprocessing is critical*. These approaches rely on comparing words in a text to a predefined dictionary of words with associated polarity or emotional scores. "Noisy" elements—such as URLs, hashtags, mentions, punctuation, and irregular capitalization—can significantly compromise the accuracy of the analysis. Therefore, preprocessing techniques like removing these

elements, normalising the text (e.g., converting to lowercase), eliminating special characters, and performing tokenization[19] and lemmatization[20] are essential for obtaining reliable results.

Furthermore, dictionary-based methods often struggle with accurately identifying complex linguistic phenomena such as irony, semantic ambiguity, and context-dependent meanings. To overcome these limitations, recent research has shifted towards pre-trained Transformer architectures, such as BERT (*Bidirectional Encoder Representations from Transformers*). BERT is a deep neural network model leveraging self-attention mechanisms to process text bidirectionally, simultaneously capturing context preceding and following each token (Devlin et al., 2019). Through pre-training on vast textual corpora with masked language modeling[21] objectives, BERT acquires rich, semantically effective contextual vector representations, excelling across numerous NLP tasks, including sentiment classification.

Compared to traditional lexical models, Transformers such as BERT are significantly more robust and less reliant on aggressive preprocessing. Being pre-trained on massive datasets from online sources, these models implicitly learn linguistic and stylistic contexts typical of the web and social media, including handling URLs, mentions, hashtags, and other informal language peculiarities. For example, BERT can recognize when a sentence is a reply because it follows a user mention, correctly interpreting the text within its conversational context. In practice, utilizing BERT for sentiment analysis involves fine-tuning: a pre-trained model is augmented with a classification layer (e.g., to label texts as positive/negative or assign numerical or star ratings) and trained on a labeled corpus, typically comprising polarized reviews, posts, or tweets. The result is a contextual classifier capable of capturing semantic nuances such as sarcasm, metaphors, or complex

[19] Tokenisation: segmenting text into individual tokens or words.

[20] Lemmatisation: reducing words to their common roots.

[21] Masked Language Modelling (MLM): A pre-training objective in which certain tokens in a text sequence are randomly replaced with a [MASK] token, and the model is trained to predict the original words based on the surrounding context. This forces the model to develop a deep understanding of word meaning and syntactic structure, enabling it to generate context-sensitive representations.

linguistic dependencies, which are challenging for purely lexicon-based methods.

Although Transformer-based models like BERT are substantially more robust against textual "noise" compared to lexical approaches, text preprocessing—such as URL removal and text normalisation—remains recommended. While cleaning is essential for models like VADER to isolate emotionally charged words, for BERT, which learns contextual representations, it ensures data consistency and computational efficiency. Pre-training on diverse, extensive corpora allows BERT to natively handle many online linguistic peculiarities, but targeted preprocessing helps focus sentiment analysis on the most semantically relevant text elements.

Numerous studies have shown Transformer-based models (such as BERT and variants like RoBERTa and DistilBERT) significantly outperform traditional lexical methods in sentiment analysis accuracy and robustness, particularly with brief and noisy texts typical of social media (Y. Liu et al., 2019; Sanh et al., 2020).

From a practical standpoint, the Python ecosystem offers several NLP libraries built on these models. Notably, Hugging Face's Transformers library (Wolf et al., 2020) provides a unified API for numerous pre-trained models, including BERT, facilitating both direct inference and custom fine-tuning. For example, you can load a pre-trained model with just a few lines of code:

```
from transformers import pipeline
sentiment_pipeline = pipeline("sentiment-analysis")
```

The resulting function directly assigns a sentiment score to a given text. In a research scenario focused on polarisation, this allows for the automated quantification of the emotional tone of thousands of posts or comments, thereby enabling the construction of sentiment score distributions for each group or community and the comparison of their differences.

Furthermore, more specialized and advanced models such as RoBERTa, XLNet, or models specifically pre-trained on social corpora like BERTweet (D. Q. Nguyen et al., 2020; Yang et al., 2020) can be

employed to further refine the analysis. This improves sensitivity to slang, emojis, and the unique linguistic characteristics of digital platforms (Yang et al. 2020).

In summary, the evolution from lexical-dictionary-based approaches (e.g., VADER) to contextual neural models (such as Transformers) represents a fundamental methodological advancement. This shift enables more accurate and nuanced measurements of emotional polarisation in texts, thanks to the power of pre-trained models and the availability of open-source Python tools that allow for the efficient implementation of large-scale sentiment analysis pipelines.

2.6 Topic Modelling

Topic Modelling represents a pivotal methodological approach for identifying, structuring, and interpreting major themes within textual content produced by digital communities. Given the substantial volume and complexity of data generated through online interactions, robust and versatile techniques capable of extracting coherent thematic structures are essential. Among the foundational probabilistic approaches is *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003), widely adopted for discovering latent semantic structures within textual corpora by modelling documents as mixtures of underlying topics. Another influential probabilistic method, *Probabilistic Latent Semantic Analysis* (PLSA), offers similar capabilities by modelling the probability distributions of words given latent topics, albeit with some computational limitations when scaling to large datasets (Hofmann, 2013).

Alternative non-probabilistic approaches such as *Non-negative Matrix Factorization* (NMF) (Lee & Seung, 1999) and classical *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990) have also demonstrated considerable effectiveness in extracting meaningful thematic structures. NMF, in particular, often yields more interpretable results due to its constraint of non-negativity, making it highly suitable for thematic extraction in social media texts. Recent developments have introduced neural and transformer-based models, notably *BERTopic* (Grootendorst, 2022) and *Top2Vec* (Angelov, 2020). These advanced techniques leverage dense contextual embeddings derived from transformer-

based models such as BERT (Devlin et al., 2019), enabling them to capture nuanced semantic relationships and significantly enhance interpretability and coherence of detected topics.

These various topic modelling techniques provide researchers with powerful tools for gaining insight into online community dynamics, enabling identification of dominant narratives, ideological clusters, and thematic divergences (Cinelli et al., 2021). Within contexts of digital polarisation, such analytical capacity is especially valuable, as it reveals specific narratives and topics driving ideological fragmentation. Topic modelling thus complements methods such as Sentiment Analysis and Social Network Analysis, supporting comprehensive understanding of both the structural and content-based dimensions of online discourse (Zhao et al., 2011; Roberts et al., 2016).

Furthermore, the integration of topic models with large language models (LLMs) has significantly enhanced analytical capacities, allowing more sophisticated inference about underlying belief systems and ideologies within communities (Grootendorst, 2022). These hybrid approaches facilitate not only descriptive identification of themes but also inferential evaluations of how specific narratives influence community structures and polarisation dynamics (Jelodar, H., Wang, Y., Yuan, C. et al., 2019; Dieng, A. B., Ruiz, F. J., & Blei, D. M., 2020).

In summary, topic modelling methodologies—ranging from classical probabilistic models like LDA and PLSA, through matrix-factorisation methods such as NMF and LSA, to advanced transformer-based neural models including BERTopic and Top2Vec—are indispensable in understanding digital communities. Their versatility and evolving sophistication continue to expand their utility, providing deeper and more precise insights into the multifaceted nature of online interactions and digital polarisation.

3. AI, LLMs, and Polarisation in Digital Communities

We currently live in the era of Artificial Intelligence (AI), a transformative age in which sophisticated computational systems

increasingly perform tasks traditionally requiring human intelligence. These tasks encompass diverse capabilities such as: learning, reasoning, problem-solving, and natural language comprehension. Among recent advances in AI, Large Language Models (LLMs) have emerged as particularly groundbreaking technologies. LLMs are deep-learning models trained on massive textual datasets, enabling them to understand, generate, and manipulate human language with remarkable precision and fluency. Prominent examples include GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers), and cutting-edge models such as *Gemini*, *Claude*, and LLaMa (*Large Language Model Meta AI*), which have become indispensable tools for processing and interpreting vast amounts of textual data generated by digital communities.

These advanced models significantly enhance the speed and accuracy of many traditional analytical approaches previously discussed, such as Sentiment Analysis and Topic modelling. Furthermore, LLMs facilitate sophisticated applications of NLP, including advanced ideological positioning techniques. Additionally, these models provide researchers with the capability to simulate polarisation dynamics within synthetic networks and implement autonomous agents for exploring complex social phenomena.

However, despite their considerable potential, LLMs are not exempt from significant limitations and concerns. Notable challenges include inherent algorithmic biases, model hallucinations stemming from excessive generative capabilities, and ethical considerations relating to fairness, transparency, and responsible application. In the following sections, these capabilities, methodological innovations, and limitations of AI and LLMs will be examined specifically within the context of analysing polarisation dynamics in digital communities.

3.1 Advanced Analytical Capabilities of LLMs in Digital Communities

Beyond the traditional applications of sentiment and topic analyses previously discussed, recent advancements in NLP, especially through LLMs, have significantly enhanced our ability to investigate nuanced dimensions of digital discourse. Particularly, advanced transformer-

based models such as BERT (Devlin et al., 2019), RoBERTa (Y. Liu et al., 2019), and GPT-based architectures (Brown et al., 2020; Radford et al., 2019) have facilitated the emergence of sophisticated analytical techniques including **ideological positioning** and **Aspect-Based Sentiment Analysis** (ABSA).

Ideological positioning allows for the precise determination of a text's or author's location along ideological or political dimensions (e.g., left-right, conservative-liberal), moving beyond traditional binary classifications and enabling a more granular representation of ideological landscapes (Le Mens & Gallego, 2025). By employing transformer-based models fine-tuned on annotated ideological datasets, researchers can accurately quantify ideological leanings from textual sources such as social media, political speeches, and online commentaries, capturing subtle rhetorical shifts and ideological clusters over time.

Complementing this, ABSA further refines traditional sentiment analysis by associating sentiment explicitly with specific aspects or topics mentioned within a text (Pontiki et al., 2014; Sun et al., 2019). Unlike conventional sentiment analysis, which assigns an overall polarity to an entire document, ABSA employs advanced NLP methods—including transformer architectures such as BERT—to detect distinct sentiment orientations toward particular entities or aspects mentioned within a sentence or a post (Hoang et al., 2019; Xu et al., 2019). Recent research has demonstrated the effectiveness of transformer-based ABSA in capturing complex sentiment dynamics within online discourse, enabling detailed insights into how specific issues or themes contribute to polarisation in digital communities (Dai et al., 2021; Zeng et al., 2019).

By integrating ABSA with ideological positioning, researchers can thus achieve a highly nuanced understanding of digital polarisation. For example, ABSA can identify precisely which aspects of political discourse evoke stronger emotional responses or polarised sentiment, while ideological positioning locates these sentiments within broader ideological spectra. This combination provides a powerful analytical framework capable of dissecting digital community interactions in unprecedented detail, significantly enhancing the explanatory depth regarding how polarisation emerges and evolves over time.

In conclusion, advanced NLP techniques, particularly ideological positioning and Aspect-Based Sentiment Analysis, substantially extend the capabilities offered by traditional sentiment and topic analyses. Leveraging the sophistication of transformer-based LLMs, these methods provide comprehensive analytical tools to reveal the finer ideological structures and nuanced sentiment dynamics that characterise contemporary digital communities.

3.2 Simulation of Polarisation Dynamics using LLMs

Recent advancements in LLMs have expanded the possibilities for simulating complex social phenomena, including polarisation, through the generation of synthetic social networks populated by autonomous agents. This innovative methodology provides an experimental framework where virtual agents, powered by LLMs, interact within simulated environments, thereby replicating human-like social dynamics and enabling controlled investigations into the mechanisms driving polarisation (Anthis et al., 2025; Argyle et al., 2023).

The construction of **synthetic social networks** leverages the generative and interactive capabilities of LLMs—such as GPT-4 and similar architectures—to model realistic social interactions at scale. In these simulations, each agent embodies distinct ideological or communicative profiles, interacting autonomously based on probabilistic and contextually coherent behaviours. Through iterative interactions, these agents spontaneously exhibit hallmark phenomena of human social networks, such as homophily (the tendency to associate with similar others), confirmation bias, and the formation of echo chambers. As interactions progress, researchers can systematically observe how specific communication patterns, information diffusion processes, and network structures influence the emergence and reinforcement of polarisation (Donkers & Ziegler, 2025)

These simulations using **autonomous LLM-based agents** represent a significant methodological leap forward. Traditional empirical approaches, such as observational studies or controlled human experiments, often face limitations in scalability, ethical constraints, and experimental control. In contrast, LLM-driven simulations allow researchers to rigorously test hypotheses regarding the causal factors of

polarisation and to explore mitigation strategies in safe, scalable, and reproducible environments (Anthis et al., 2025; Chuang et al., 2024). By adjusting parameters such as ideological orientation, susceptibility to influence, or network connectivity, scholars can systematically analyse how varying conditions impact polarisation dynamics.

Moreover, LLM-based agent simulations provide valuable insights into potential strategies for polarisation mitigation. For example, by systematically varying agents' openness to diverse viewpoints or exposure to counter-attitudinal information, researchers can identify conditions under which polarisation can be effectively reduced or exacerbated, providing crucial guidance for real-world interventions (Donkers & Ziegler, 2025; Wang et al., 2025).

In conclusion, synthetic social networks powered by autonomous LLM-driven agents constitute a powerful experimental paradigm for studying the dynamics of polarisation. This approach offers unprecedented opportunities to observe, analyse, and intervene in complex social phenomena at a level of detail and scale unattainable through traditional methods, significantly enhancing our understanding of polarisation in digital and social contexts (Piao et al., 2025).

4. Critical Reflections on the Various Methodologies

Before addressing the critical issues associated with traditional and emerging approaches, Table 3 summarises the key methodologies and metrics discussed in the previous sections.

Table 3 - Statistical Methodologies for Analysing Polarisation

Method.	Metric/Index	Primary Purpose	Brief Interpretation
SNA	Centrality (*Degree, Closeness, Betweenness*)	Measures the structural importance and influence of individual users (nodes) within a network.	*Degree*: direct popularity. *Closeness*: efficiency in spreading information. *Betweenness*: bridging role between distinct groups.
SNA	Modularity (Q)	Quantifies the strength of division within a network into distinct communities (echo chambers).	$Q > 0.3$ indicates a strong and meaningful community structure, suggesting network polarisation.
SAD	Hartigan's Dip Test (*p*-value)	Provides a rigorous statistical test to assess whether a distribution of opinions is unimodal (consensus) or multimodal (polarisation).	A very low p-value (e.g., < 0.05) leads to the rejection of unimodality, confirming the presence of multiple opinion peaks.
	Sarle's Bimodality Coefficient (*b*)	Combines skewness and kurtosis into a single index to detect bimodality tendencies.	Values of $b > 5/9 \approx 0.555$ suggest potential bimodality, and thus polarisation.
	Esteban & Ray Polarisation Index	Measures polarisation by accounting for both the distance between opinion groups and their sizes.	Index is maximised when the population is evenly split into two groups with divergent positions.
	Extreme Polarisation Ratio	Provides a simple and direct measure of the share of ratings falling at the extremes of a scale.	Values close to 1 indicate that nearly all opinions are extreme (e.g., 1 or 5 stars), signalling high polarisation.
SA	Sentiment Score (e.g., VADER, BERT, etc.)	Measures the orientation (positive/negative) and emotional tone intensity of textual content.	The distribution of scores within a group reveals emotional valence; divergence between group scores reflects affective polarisation.
ABSA	Aspect-based Sentiment Analysis	Refines SA by associating sentiment explicitly with specific aspects mentioned within a text	Enabling detailed insights into how specific issues contribute to polarisation in digital communities.
TM	Topic Modeling (e.g., LDA, LSA)	Identifies latent topics in textual corpora to detect thematic clustering and divergence across groups.	Separate topic distributions between groups may indicate fragmentation or echo chambers.

4.1 Limitations of Traditional Approaches

Despite their analytical power, traditional quantitative methodologies are not without limitations. It is essential for researchers to be aware of these constraints in order to interpret results correctly and with nuance. As argued throughout this chapter, a holistic approach is indispensable precisely because the weaknesses of one method can be partially offset by the strengths of another.

Limitations of Social Network Analysis (SNA).

- *Static vs Dynamic Nature:* Standard SNA provides a "snapshot" of the network at a specific point in time. This static approach may fail to capture the temporal evolution of communities and polarisation dynamics, which are fluid and constantly changing.
- *Structure Without Context:* While SNA effectively reveals the structure of interactions—who communicates with whom—it does not explain *why* or *about what*. A network may appear polarised, but only content analysis can determine whether the division concerns a specific issue or stems from broader social dynamics. For this reason, SNA alone is insufficient and must be integrated with sentiment and distributional analyses to verify whether structural divisions correspond to actual opinion polarisation.
- *The Invisibility of "Lurkers":* Networks are constructed based on visible interactions (e.g., retweets, replies, mentions). This excludes the large population of passive users, or "lurkers", who read and are exposed to content without interacting. Their role in the polarisation ecosystem remains invisible to this methodology (Nonnecke & Preece, 2000).

Limitations of Sentiment Analysis (SA).

- *Sarcasm and Cultural Context:* Although advanced models such as BERT perform well, understanding complex nuances—such as irony, sarcasm, or community-specific

jargon—remains a significant challenge. A statement may be classified as positive based on word choice, even if it is intended sarcastically.

- *Limits of Text-Based Analysis:* Online communication is increasingly multimodal (including images, memes, videos, emojis). Sentiment analysis based solely on text cannot capture the emotional tone or meaning conveyed by these elements, which are often crucial for interpreting the message.
- *Sentiment as a Proxy:* It is important to remember that sentiment scores are proxies for expressed emotion, not direct measures of an individual's psychological state. People choose their language strategically, and what they write may not fully reflect how they actually feel.

Limitations of Statistical Analysis of Distributions (SAD).

- *Explaining the Shape:* Much like SNA, distributional analysis describes *what* is observed, but not *why*. A bimodal distribution of reviews may reveal the existence of two groups—"lovers" and "haters"—but it does not explain the reasons behind this divergence (e.g., product flaws, unmet expectations, political alignment).
- *Bimodality Is Not Always Polarisation:* Detecting two peaks in a distribution is a strong indicator of polarisation, but not definitive proof. It may result from data artefacts or from the presence of two distinct populations that are not necessarily in conflict (e.g., novice vs expert reviewers).
- *Sample Size Sensitivity:* Statistical tests such as the Dip Test require a sufficiently large sample size to be reliable. In small samples, it may not be possible to reject the null hypothesis of unimodality—even if polarisation is present in the broader population.

It is worth noting that these limitations do not invalidate the methodologies presented, but rather reinforce the chapter's central argument: only a holistic, multi-method approach can adequately capture the complexity of digital polarisation.

4.2 Limitations, Challenges, and Ethical Considerations on LLMs

Despite their remarkable analytical capabilities, LLMs present significant methodological, technical, and ethical challenges. A primary concern is the intrinsic *algorithmic bias* embedded within these models, originating largely from training datasets that reflect existing societal biases and inequalities (Bender et al., 2021; Ferrara, 2023). As a result, LLMs can unintentionally perpetuate stereotypes, political biases, and discriminatory narratives when generating or interpreting content, thereby inadvertently exacerbating existing societal divides (Weidinger et al., 2021).

A related challenge is the phenomenon known as "*hallucination*", whereby LLMs produce factually incorrect or nonsensical outputs that appear plausible and coherent (Ji et al., 2023). Such hallucinations can substantially undermine the reliability and validity of LLM-generated analyses, especially when utilised in sensitive contexts like political discourse or policy-related communication. Furthermore, the effectiveness of these models critically depends on data quality. Erroneous, biased, or low-quality training data can propagate misinformation, distort sentiment analysis outcomes, and skew ideological positioning results, thus compromising the accuracy and ethical soundness of analytical insights (Ferrara, 2023; Ji et al., 2023).

Another significant ethical issue concerns the potential for LLMs to induce or amplify polarisation. Because these models are adept at replicating persuasive and ideologically congruent communication, they can reinforce echo chambers or inadvertently escalate ideological divisions within digital communities. Therefore, researchers must exercise caution and rigorously validate LLM-derived insights to ensure responsible deployment and mitigate unintended societal impacts.

The future application of AI and LLMs in studying digital polarisation appears promising but necessitates further innovation and methodological refinement. **Multimodal approaches**, which integrate textual analysis with audio, visual, and behavioural data, present a promising pathway to achieve a more comprehensive understanding of social dynamics. Recent advancements in multimodal AI models can capture subtle emotional and non-verbal signals within online interactions, offering richer, contextually nuanced insights into how polarisation

emerges and evolves (Al-Shehri & Al-Qahtani, 2025; Durrheim & Quayle, 2025).

Furthermore, there is an urgent need for an **interdisciplinary approach** to ensure the responsible and ethical development of AI within social sciences. Collaboration between computer scientists, social scientists, psychologists, ethicists, and policymakers is crucial to comprehensively address the technical complexities, societal implications, and ethical considerations inherent in deploying LLM-based analytical tools. Such collaborative frameworks can foster the development of ethically-aligned AI methodologies, rigorous validation procedures, and clear standards for data governance and transparency, ultimately enabling the beneficial and socially responsible use of LLMs in polarisation research (Thapa et al., 2025).

In conclusion, while AI and LLMs offer powerful tools for deepening our understanding of digital polarisation, realising their full potential requires addressing their inherent limitations through multidisciplinary collaboration, methodological innovation, and vigilant attention to ethical standards.

Part III

PERSONALITY AND DIGITAL COMMUNITIES

Understanding personality traits from written language represents a crucial step in the analysis of complex social phenomena, including polarisation in digital communities.

Personality traits significantly influence the ways in which individuals interact online, interpret information, and engage in public discussions. For instance, high *Openness* to experience may predispose users to engage with divergent opinions and heterogeneous content, whereas high *Neuroticism* may make them more susceptible to emotionally charged messages, increasing the likelihood of polarised reactions.

The ability to automatically infer personality from text thus enables not only the study of individual characteristics but also the modelling and prediction of collective dynamics such as the amplification of ideological conflicts or the formation of echo chambers.

This chapter begins by presenting the principal psychometric models for personality prediction, such as the *Big Five* and the *Myers–Briggs Type Indicator* (MBTI). It then examines personality prediction techniques, ranging from the pioneering lexical approaches of Pennebaker to classical machine learning methods and, more recently, advanced approaches based on Transformer models. In this perspective, personality measurement through textual analysis can serve as an explanatory variable within models aimed at accounting for polarisation in digital communities.

1. Measuring Personality through Text Analysis

1.1 Lexical Hypothesis and Psychometric Models: A Comparison between the Big Five and the MBTI

Text-based personality measurement rests on a cornerstone of personality psychology: the lexical hypothesis. This hypothesis posits that the most relevant individual differences tend to be encoded in natural language (Allport & Odbert, 1936; Goldberg, 1990). In other words, languages develop terms over time to describe the most important personality traits. This premise has led to the construction of various psychometric models. Among them, two frameworks are especially prominent in computational applications, albeit with very different scientific standing: the Big Five model (Goldberg, 1990) and the Myers–Briggs Type Indicator (MBTI) (Myers, 2003).

The **Big Five** is the dominant and most thoroughly validated framework in personality psychology. It is a dimensional model that describes personality through five core traits: *Openness* to Experience (Openness), *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*. The robustness of this model is supported by a large corpus of empirical evidence and by the cross-cultural replicability of the five factors, making it the gold standard in the scientific assessment of personality (Shum et al., 2025). Moreover, the Big Five provide measurements on continuous scales, enabling a nuanced description of individual differences rather than a rigid typological classification.

The **MBTI** is a psychometric instrument grounded in Jungian typology; the profile is derived from a self-report questionnaire and classifies individuals into 16 discrete types based on four binary dichotomies: Introversion (I) vs Extraversion (E), Sensing (S) vs Intuition (N), Thinking (T) vs Feeling (F), and Judging (J) vs Perceiving (P). Despite its widespread use in organisational and personal-development contexts, reliability estimates depend on the metric employed. Test–retest correlations for the continuous preference scales (E–I, S–N, T–F, J–P) are generally good in recent versions (MBTI Step I), with coefficients in the range 0.81–0.86 over intervals of 6–15 weeks (Myers-Briggs Foundation, 2025). By contrast, the stability of the four-letter "type" is lower: historical studies on earlier forms report that roughly

half of respondents change at least one preference after about five weeks (Howes & Carskadon, 1979; Pittenger, 2005). These differences reflect both the effects of dichotomisation (unstable borderlines) and the evolution of test forms. In summary, while the MBTI is broad and popular in applied settings, it presents psychometric limitations when interpreted as a rigid typology; it performs more reliably when treated as a set of preference scales.

A direct comparison between the two models highlights the superior scientific merits of the Big Five. The latter provide richer and more reliable information on individual traits than the MBTI and exhibit robust external correlates and predictors (Celli & Lepri, 2018). Some MBTI dimensions show partial correspondences with the Big Five—for example, the MBTI Introversion/Extraversion dichotomy correlates with the Extraversion trait in the Big Five—but overall the Big Five offer a more accurate and comprehensive representation of personality. Other correspondences, identified heuristically between the two models, are presented in Table 4, which also indicates, for each trait, typical behaviours/tendencies associated with high and low scores.

It has been observed that the MBTI's enduring popularity stems in part from its positive and engaging language: it avoids labelling traits with negative connotations (the MBTI does not include an explicit analogue of Neuroticism), which makes it more acceptable to the general public than measures that incorporate less flattering aspects of personality (Miller & Lovler, 2018). However, this accessibility comes at the expense of scientific rigour.

In natural language processing applied to personality, a critical paradox emerges concerning data availability. The psychometrically superior model—the Big Five—is paradoxically more difficult to apply at scale because of the scarcity of labelled data, whereas the weaker model—the MBTI—is extremely easy to obtain online. In numerous forums and social platforms, many users spontaneously disclose their MBTI type (the four letters) in profiles or discussions, providing a rich source of self-labelled data that can be automatically collected in large quantities. By contrast, individuals' Big Five scores are rarely stated explicitly online; to obtain them, users must have completed a validated

personality test and shared the results. Retrieving Big Five data therefore requires more complex strategies—for example, searching for posts in which users report their scores or percentiles—often combined with semi-automatic extraction and manual verification (Shum et al., 2025).

Table 4 – Principal personality models and traits with behavioural description

Trait	**Models**	**High score focus**	**Low score focus**
Openness	Big Five; MBTI (N – S)	Curious Imaginative Creative Open to new ideas Intuitive	Practical Routine-oriented Conservative in thinking Sensing-focused
Conscientiousness	Big Five; MBTI (J – P)	Organised Responsible Goal-oriente Judging	Flexible Spontaneous Less structured Perceiving
Extraversion	Big Five; MBTI (E – I)	Energetic Sociable Enthusiastic Extrovert	Reserved Introspective Prefers solitude Introvert
Agreeableness	Big Five; MBTI (F – T, *partial overlap*)	Compassionate Cooperative Trustworthy Feeling	Competitive Sceptical, Less accommodating Thinking
Neuroticism	Big Five; MBTI (*no direct analogue*)	Emotionally Reactive Prone to stress, Mood variability	Calm Emotionally stable Resilient Stress-tolerant

Note: MBTI correspondences are heuristic analogues (not 1:1 mappings); Neuroticism has no direct MBTI axis.

This disparity has introduced a convenience bias into the datasets available for research: many recent computational studies on personality use the MBTI simply because it is easy to obtain large quantities

of labelled data in that format (Celli & Lepri, 2018; Saeteros et al., 2025). This entails a serious methodological risk linked to self-reference bias in MBTI-based analyses. Machine-learning models trained on data in which users openly disclose their type tend to 'learn' spurious shortcuts: for example, the presence of the string 'INFP' in a text becomes a strong cue for predicting the INFP type—not because there are subtle linguistic markers signalling that trait, but simply because the user has explicitly mentioned their type. Consequently, an MBTI classifier may achieve high accuracy by picking up such self-disclosures rather than genuine linguistic signals of personality. Targeted studies have confirmed the problem: when all explicit references to types (the four-letter codes) are masked in the text, the performance of previously strong MBTI models drops sharply, revealing the extent to which accuracy scores had been inflated by those trivial cues (Saeteros et al., 2025). Therefore, findings from research based on self-declared MBTI datasets should be interpreted with extreme caution. This scenario underscores the urgency of developing large-scale corpora with reliable Big Five labels, or devising techniques that compensate for the intrinsic biases of existing MBTI data, thereby realigning research with psychometrically stronger personality measures.

1.2 From Lexicons to Vectors: Early Computational Approaches to Personality Recognition

Before the advent of deep learning, automatic personality recognition from text developed along two main lines: methods based on predefined psychological lexica and classical machine-learning models with engineered features.

***Lexicon-based* methods (e.g., LIWC).**

A pioneering approach to linking language and personality relied on psychological dictionaries, the best known of which is LIWC (Linguistic Inquiry and Word Count) by James Pennebaker and colleagues. LIWC operates by counting, within a text, the frequency of terms belonging to predetermined lexical categories—such as pronouns, positive or negative emotion words, and terms referring to cognitive processes—that have psychological relevance (Tausczik & Pennebaker,

2010). The underlying idea is that the way people use certain classes of words—especially function words (pronouns, articles, prepositions)—reveals cues about their psychological states and personality traits. This premise is supported by compelling findings. For example, an analysis with LIWC showed that frequent use of first-person singular pronouns ("I", "me") is not in itself an indicator of narcissism; rather, it can correlate with heightened self-awareness and, in some contexts, with depressive states or subordinate social roles (Pennebaker, 2011). Likewise, large-scale studies suggest that more extraverted individuals tend to employ language richer in positive emotions and social references (Koutsoumpis et al., 2022), whereas individuals high in *Neuroticism* more often use expressions associated with negative affect or anxiety. Although lexicon-based methods are simple and interpretable, their effectiveness depends on dictionary coverage: they can only detect what is defined a priori in the categories and may miss important linguistic nuances not captured by the lexicon.

"Classical" machine-learning models.

With the growth of social media and the availability of large sets of labelled texts (for example, via personality tests completed by users), research shifted towards supervised machine-learning methods. In this approach, numerous quantitative features are automatically extracted from text and then fed into classification or regression algorithms to predict personality traits.

During this period, the algorithms most frequently employed included: Support Vector Machines (SVM), Random Forests (RF), Naive Bayes (NB), Gradient Boosting (GB), and Logistic Regression (LR) (Al-Fallooji & Al-Azawei, 2022). When properly trained, these algorithms were applied to both classification problems (e.g., predicting a person's MBTI type) and regression problems (predicting continuous Big Five scores).

The success of these models largely depended on choosing the right features. Hundreds of indicators extracted from each textual profile were often combined: word-frequency vectors with TF-IDF weighting, scores across LIWC categories, sentiment metrics (affective valence of the text), part-of-speech distributions (proportions of nouns, verbs, adjectives, etc.), readability measures, and even

behavioural metadata such as posting frequency or posting time (Safitri & Setiawan, 2022; Bhangale, 2025). The idea was to characterise a user's communicative style as comprehensively as possible, providing the learning algorithm with the widest set of cues correlated with personality.

These classical approaches demonstrated the feasibility of automatic personality prediction, albeit with accuracies and performance that are modest by today's standards. For example, Safitri and Setiawan (2022) report 88% accuracy in classifying Twitter users on the five Big Five traits using an SVM classifier optimised on their tweets. Other studies focusing on the MBTI found that a Random Forest could reach about 77% accuracy in predicting personality type from a user's social posts (Bhangale, 2025), while Stracqualursi & Agati (2025) predicted YouTube users' personality with an XGBoost model, showing that reliable prediction was achievable with as few as 100 words. It should be noted, however, that performance varied substantially depending on the model employed, the type of features extracted, and the size/quality of the dataset; the literature also reports more modest results on other corpora, especially for traits that are harder to capture.

This technological evolution underscores a growing trade-off between model interpretability and predictive power. Lexicon-based approaches such as LIWC offer maximal transparency: the output is essentially a count of words in psychologically motivated categories. This enables researchers to draw direct conclusions—for example, "*in texts produced by people with high Extraversion, one observes more frequent use of positive-emotion terms and social references*" (Koutsoumpis et al., 2022). By contrast, classical statistical models such as SVMs or Random Forests increase accuracy but partially obscure the reasons behind predictions. Some features remain readily interpretable (e.g., a high sentiment score associated with high Extraversion is intuitive), yet the decision boundary of an SVM in a multidimensional space of thousands of TF-IDF terms is far more difficult to interpret. Interpretability in these cases is often recovered ex post, for instance by analysing feature importance within the model (Bhangale et al., 2025) or by observing how predictions change when certain cues are removed from the text.

The advent of deep neural models, particularly Transformers, pushed this trade-off further: these architectures deliver unprecedented predictive performance but are often considered "black boxes" (Bama et al., 2025). They operate with millions (if not billions) of parameters and highly non-linear interactions, making a direct understanding of why a given text is associated with a given personality score prohibitively complex. This trend has motivated the development of explainable-AI (XAI) techniques applied to personality computing, in an effort to recover at least part of the model's readability. For example, recent studies have employed attribution methods such as Integrated Gradients (Sundararajan et al., 2017) to highlight which words in a text contribute most to a Transformer's prediction (Devlin et al., 2019). Such analyses have shown, for instance, that in a text judged highly neurotic by the model, terms like "worried" or "anxious" receive high importance weights, providing a qualitative indication consistent with psychological theory. Approaches of this kind represent an attempt to reconcile performance and interpretability, enabling researchers to assess whether the decision basis of a complex model is plausible (for example, whether, to predict Agreeableness, the model relies on polite or prosocial words rather than unrelated elements). This line of work on interpretability is now a fundamental component of the field, to ensure that gains in accuracy do not come at the expense of human understanding and trust in personality-computing systems.

1.3 Contextual Embeddings and Bidirectional Architectures

The introduction of Transformer architectures has revolutionised natural language processing and, consequently, applications in personality prediction. Models such as BERT (*Bidirectional Encoder Representations from Transformers*; Devlin et al., 2019) and its more advanced variants, e.g. RoBERTa (Y. Liu et al., 2019), marked a shift from representations based on predefined features to representations automatically learned from text. The key innovation of these models lies in their ability to produce dense, contextual embeddings: each word in a text is transformed into a numerical vector that captures its meaning in relation to the surrounding context, enabled by the self-attention mechanism at the core of Transformers (Vaswani et al., 2017).

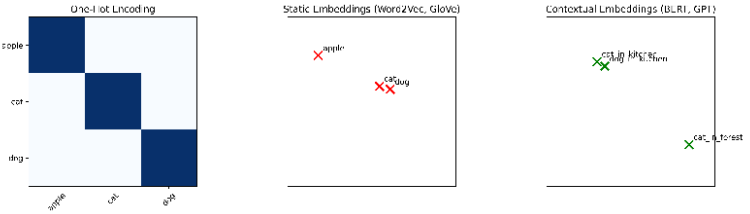

Figure 5. Differences between One-Hot Encoding, Static and Contex-tual Embeddings

Figure 5 illustrates how word representations evolve from a discrete and rigid form (one-hot encoding) to denser, semantic embeddings, and ultimately to contextual ones.

In personality prediction, the typical procedure for applying these models is *fine-tuning*. The process begins with a model pre-trained on a very large corpus of general text (e.g. Wikipedia, books, public forums), which already encodes a basic linguistic knowledge. The model is then re-trained on a dataset specifically labelled with the personality traits of interest (Shum et al., 2025). During fine-tuning, a small neural layer (often a simple linear classifier) is added on top of the Transformer to generate the final prediction—for example, the probabilities of the 16 MBTI classes, or five numerical values estimating an individual's Big Five scores. Training adjusts the parameters of the model so that, given a user's text collection as input, the output of the final layer corresponds to that user's known personality labels.

A large body of comparative studies has demonstrated the performance leap achieved with Transformers over previous approaches. Optimised models such as RoBERTa (Y. Liu et al., 2019)—essentially an improved version of BERT trained on larger datasets and with refined training procedures—tend to outperform baseline BERT in personality prediction tasks. For example, RoBERTa has shown higher accuracy and lower mean squared error than BERT in predicting Big Five scores from users' posts under identical experimental conditions. Furthermore, larger variants (e.g. RoBERTa-large, with hundreds of millions of parameters) often achieve superior results, provided that the fine-tuning dataset is sufficiently large to avoid overfitting and leverage their greater expressive capacity.

Another advantage of Transformers is their inherently **bidirectional** nature (in models such as BERT and RoBERTa). Unlike previous sequential models (e.g. LSTMs or models with static embeddings such as word2vec), a bidirectional model can exploit both the left and right context of a word to determine its representation. This is particularly valuable for capturing semantic nuances: for instance, the word *warm* will have different embeddings in the phrases *"a warm and welcoming character"* versus *"an oppressive warm heat"*, enabling the model to distinguish between affective and descriptive contexts relevant to personality.

In summary, the advent of BERT, RoBERTa, and their successors has provided researchers with an extraordinary tool: models that learn directly from text which linguistic features matter, without requiring humans to predefine lexica or rules. This not only increases predictive accuracy but also opens the door to qualitative analysis of latent features. By examining the embeddings or attention weights of a Transformer fine-tuned on personality, researchers can explore which words or sentences are considered salient by the model, potentially yielding new insights into the links between language and personality traits.

1.4 Taxonomy, Operational Pipeline, and Data

Recent reviews of the literature on personality prediction models have proposed a unified taxonomy that distinguishes three main families (Fig. 6):

(1) traditional machine learning models (e.g., SVM, Random Forest, Naïve Bayes), commonly referred to as *shallow learning*, which do not learn hierarchical feature representations from data as deep networks do;
(2) *sequential deep learning* models, a subfamily of deep learning specifically designed to process textual data in sequential form and to capture temporal dependencies (e.g., RNN, LSTM, GRU);
(3) *Transformer-based* models, which, through the self-attention mechanism, are able to capture far more complex and long-range contextual relationships compared to sequential models.

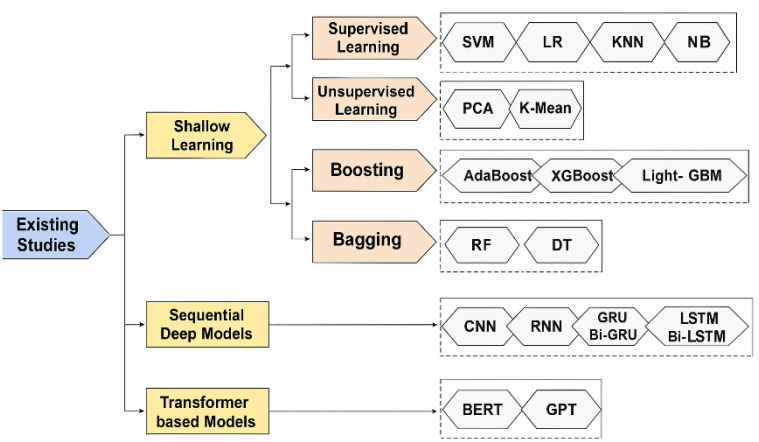

Figure 6. Taxonomy of actually existing studies

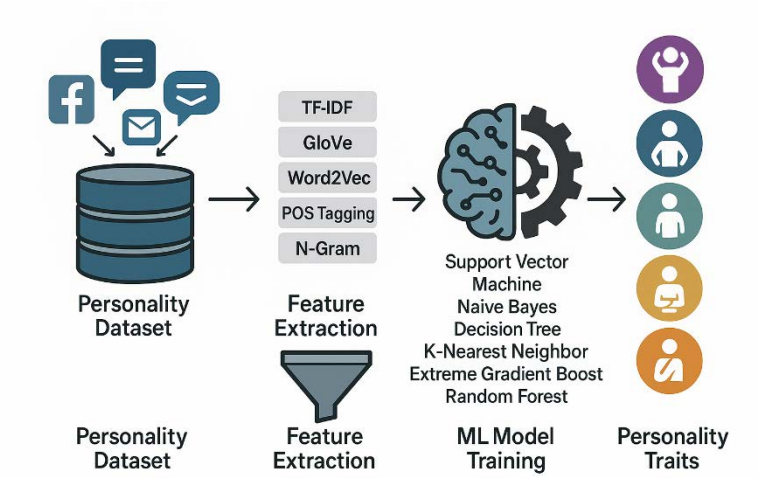

Figure 7. Basic steps in ML process for personality trait prediction

Figure 7 schematically illustrates a typical operational pipeline for predicting personality traits from textual data. The initial phase consists of collecting and preparing a labelled personality dataset (see Table 5). The texts are then transformed into numerical representations (*feature extraction*) through several techniques, including:

- **TF-IDF**, a statistical measure that reflects the importance of a term within a document relative to the corpus.
- **Word embeddings** (e.g., Word2Vec, GloVe), dense representations that capture semantic and syntactic relationships between words.
- **POS tagging**, the annotation of parts of speech, useful for identifying grammatical patterns.
- **N-grams**, sequences of *n* words used to represent co-occurrence patterns.

The extracted features are subsequently used to train various machine learning models, including Support Vector Machines (SVM), Naïve Bayes, Decision Trees, K-Nearest Neighbors, Extreme Gradient Boosting (XGBoost), and Random Forests. The final output consists of the classification or estimation of an individual's personality traits (e.g., MBTI or Big Five).

Once trained and validated, the model can be applied to new textual data for personality trait classification. During training, it is standard practice to partition the dataset into three subsets — training, validation, and test — in order to optimize hyperparameters and provide an unbiased evaluation of model performance.

Table 5 reports the most widely used English-language resources for text-based personality prediction, covering both MBTI and Big Five measures, together with their content and access links. These corpora typically contain textual material (tweets, essays, dialogues) and, in some cases, additional demographic information.

All datasets are released in anonymised form to remove personally identifiable information. However, labelling practices vary across resources: some provide annotations curated by experts or clinically supervised, whereas many rely on self-reported assessments (e.g., MBTI types or Big Five scores) supplied by participants. In any case, it is essential to consult the documentation of each dataset in order to verify the labelling protocol and known sources of bias. Expert involvement may enhance reliability, but it does not eliminate potential bias in the labels.

Table 5 – Personality Prediction Datasets

Dataset	**Attributes**	**Type**	**Sources**	**availability**
MBTI	50 tweets per user with self-declared MBTI labels (Personality Café Forum), for a total of 8,600 posts.	MBTI	https://www.kaggle.com/datasets/datasnaek/mbti-type	Public
NeoNyx	Personality typeprofile description, along voting about 35k profiles	MBTI	https://www.kaggle.com/datasets/abdelkarimbsalah/nyx-2-1	Public
Pandora	Reddit comments labeled with personality models and demographics for more than 10k users	Hybrid (MBTI, Big Five)	https://psy.takelab.fer.hr/datasets/all/pandora/	Upon request
Big Five Personality	1,015,342 questionnaire answers collected online by Open Psychometrics.[22]	Big Five	https://www.kaggle.com/datasets/tunguz/big-five-personality-test	Public
Essays I	2,468 anonymous essays with the author's self-reported personality traits[23]	Big Five	https://github.com/SenticNet/personality-detection	Public

[22] https://openpsychometrics.org/tests/IPIP-BFFM/

[23] The "Essays I" dataset, created by Pennebaker and King (1999), is a collection of 2,468 anonymous essays, each tagged with the author's self-reported Big-

2. Predictive Effectiveness and Trait-Specific Challenges

2.1 Predictive effectiveness

The evaluation of personality computing models framed as classification tasks relies on a core set of standard metrics: accuracy, precision, recall, F1, and AUC-ROC. The choice of metric must take into account the characteristics of the dataset (e.g. class imbalance, number of classes) as well as the specific objectives of the NLP analysis. The review presented here summarises and formalises these metrics, while also providing the rationale for their use in the context of text-based personality prediction.

The confusion matrix is a compact tabular representation of a classifier's performance, comparing true labels against predicted ones. For a binary problem (two classes: positive/negative), its structure is shown in Figure 8.

		Predicted	
		Predicted Positive	**Predicted Negative**
Actual	**Actual Positive**	True Positive (TP)	False Negative (FN)
	Actual Negative	False Positive (FP)	True Negative (TN)

Figure 8. Confusion matrix for a binary classification problem

Where:

- TP (True Positives): cases predicted as positive that are actually positive.
- TN (True Negatives): cases predicted as negative that are actually negative.

Five personality traits. Recently used and available on GhitHub (Majumder et al., 2017).

- FP (False Positives): cases predicted as positive but actually negative (false alarm).
- FN (False Negatives): cases predicted as negative but actually positive (missed detection).

All classification metrics are derived from these four fundamental quantities (TP, TN, FP, FN). The most commonly reported include:

- **Accuracy:** the proportion of correctly classified cases over the total number of cases.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{10}$$

This measure is intuitive and easy to interpret: for example, if a model correctly classifies 90 out of 100 cases, its accuracy is 90%. However, it may be misleading when classes are imbalanced, as performance can be dominated by the majority class. It should therefore be complemented by metrics that are more sensitive to the distribution of errors.

- **Precision:** the correctness of positive predictions.

$$Precision = \frac{TP}{TP + FP}, \tag{11}$$

This metric is particularly relevant when the cost of false positives is high (e.g. attributing the trait of Extraversion to an individual who does not actually possess it).

- **Recall:** also known as the True Positive Rate (TPR) or *sensitivity*, it is the proportion of actual positive cases that the model correctly identifies.

$$Recall = \frac{TP}{TP + FN}, \tag{12}$$

This becomes crucial when it is costly to "miss" positive cases (e.g. failing to detect individuals who truly display a given trait).

- **F1-score:** the harmonic mean of *precision* and *recall*; this metric balances the two types of error, namely false positives (FP) and false negatives (FN).

$$\text{F1-}score = \frac{2(Precision * Recall)}{Precision + Recall}, \tag{13}$$

The F1-score penalises models that perform well on one metric but poorly on the other. For example, a model with perfect precision (no false positives) but very low recall (many false negatives) will not achieve a high F1-score. For this reason, the F1 metric is particularly useful in imbalanced-class settings, where accuracy alone can be deceptive, as well as in multi-class classification problems. In these cases, it is advisable to report both **macro-F1** (the average of F1 values across classes, assigning equal weight to each) [24] and **micro-F1** (computed by globally aggregating TP, FP, FN, thus giving greater weight to larger classes) [25].

- **ROC and AUC:** The ROC curve (*Receiver Operating Characteristic*) is a graphical tool for evaluating the performance of a classification model, by plotting the True Positive Rate (TPR, also referred to as *recall* or *sensitivity*) against the False Positive Rate (FPR) [26]. The curve illustrates how these two quantities vary as the decision threshold is adjusted.

---

[24] Macro-F1 (Class-wise Mean): computes the F1-score separately for each class (e.g. introvert, extravert, Judging, etc.), and then takes the simple arithmetic mean. Macro-F1 assigns equal weight to each class, regardless of the number of instances belonging to it in the dataset. In imbalanced datasets, where a minority class may be of particular importance, macro-F1 reveals whether the model is able to classify it accurately, without its performance being "hidden" by the excellent results on the majority class.

[25] Micro-F1 (Global Aggregate): aggregates true positives (TP), false positives (FP), and false negatives (FN) across all classes, and then computes a single F1-score. Micro-F1 assigns greater weight to larger classes. Essentially, it evaluates the overall performance of the model on the entire dataset. It is very similar to accuracy, particularly in imbalanced-class settings, and indicates the model's general ability to make correct predictions.

[26] With TPR defined as in (12) and FPR = FP / (FP + TN).

The decision threshold represents the probability value above which the model assigns an instance to the positive class. For example, a classifier may decide that an individual is "extraverted" only if the estimated probability exceeds 0.8, whereas with a more permissive threshold a value of 0.7 might suffice. The choice of threshold therefore determines a different balance between false positives and false negatives.

The **AUC** provides a global measure of discriminative ability (0.5 ≈ random; 1.0 ≈ excellent). A common estimate of AUC uses the trapezoidal rule.

$$AUC \approx \sum_{I=1}^{n-1} \frac{(FPR_{i+1} - FPR_i)(TPR_{i+1} + TPR_i)}{2}, \tag{14}$$

In dichotomised trait prediction (e.g. Introvert vs Extravert), an AUC close to 1 indicates strong class separation, whereas values near 0.5 indicate performance no better than chance.

In general, accuracy is the most frequently reported metric; however, as already noted, it can be misleading in the presence of class imbalance. In such cases, **macro-F1** (or micro-F1 where imbalance is less critical) and AUC-ROC are recommended for a more reliable evaluation.

Naz et al. (2025) emphasise the importance of a "minimum reporting standard" including:

- accuracy and F1 for each trait,
- the confusion matrix,
- stratified cross-validation,
- separation of users across train/test sets to avoid data leakage,
- full documentation of preprocessing steps.

It should be noted that much of the literature, including Naz et al. (2025), focuses on trait classification (e.g. MBTI or binary thresholds for the Big Five), while neglecting regression metrics such as RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), or R^2 (Coefficient of Determination). Yet, when personality scores are available in continuous form, the use of these regression metrics is necessary for

a proper evaluation of models. The choice of metric should therefore always be aligned with the nature of the prediction target.

2.2 Predictive Challenges for Specific Traits

A consistent finding across several studies is that not all personality traits are equally predictable from text, both in the MBTI framework (Choong & Varathan, 2021; Naz et al., 2025) and in the Big Five model (Ramon et al., 2021; Stewart et al., 2022). Within MBTI, one of the most difficult traits to predict is **Judging vs Perceiving**, for which models generally achieve only marginal levels of accuracy. Choong & Varathan (2021), in a systematic review of major studies, highlight the inherent challenge of predicting this dichotomy.

With respect to the Big Five traits, results can be broadly summarised as follows:

- **Easier-to-predict traits.** *Openness to Experience* and *Extraversion* are typically the most predictable. Openness often yields the lowest prediction error (i.e. the model's estimates come closest to actual values), while Extraversion tends to achieve the highest R^2 coefficients or the best binary classification accuracy. This suggests that the linguistic signal for these traits is relatively strong and consistent.
- **Harder-to-predict traits.** *Conscientiousness*, *Agreeableness*, and *Neuroticism* present significant challenges. Models consistently struggle to achieve satisfactory performance: explained variance (R^2) for these traits is often very low (frequently close to zero), and estimation errors remain high (Shum et al., 2025). Even in binary classification tasks (e.g. distinguishing between high and low Conscientiousness), performance is generally only slightly above chance.

Another important observation is that the five traits are not fully independent but display well-known *intercorrelations* in personality psychology (McCrae & Costa, 1987). For instance, highly open individuals are somewhat more likely to be extraverted, while *Extraversion* and *Agreeableness* may also correlate positively. Exploiting these interrelations can enhance prediction: multi-task learning approaches—in

which a single model is trained to predict all five traits simultaneously—have consistently outperformed training separate models for each trait. In practice, a Transformer with a multi-task architecture learns internal representations of a user's text that encode information reusable across traits, capitalising on their interconnections. For example, strong signals of *Extraversion* and *Openness* in a user's text may moderately and coherently influence the estimation of other traits.

Why are some traits so much harder to predict? The prevailing explanation is that this reflects genuine differences in the degree to which traits manifest themselves linguistically (Golbeck et al., 2011; Schwartz et al., 2013). Traits such as Openness and Extraversion have relatively direct and recognisable verbal markers. Highly open individuals are more likely to refer to artistic interests, abstract ideas, travel, or culture, and employ varied, complex vocabulary. Highly extraverted individuals frequently use socially engaging words (friends' names, nicknames), emoticons, or laughter, and express positive emotions such as enthusiasm and joy. These are surface-level signals that emerge clearly in text.

By contrast, traits such as *Conscientiousness* or *Agreeableness* are more closely tied to behaviours and attitudes not always verbalised. Conscientiousness involves organisation, reliability, and diligence—qualities that may surface only indirectly in text (e.g. avoidance of profanity, structured language) but are more clearly expressed through actions (meeting deadlines, keeping promises). Similarly, *Agreeableness* (friendliness, empathy, cooperativeness) is often manifested through everyday prosocial behaviours, which may not be readily evident in written communication on social media. *Neuroticism* (tendency towards anxiety and emotional instability) may even be deliberately concealed: as users are aware that continuous complaints or displays of vulnerability may be viewed negatively, highly neurotic individuals may intentionally mask this trait in public posts—an example of social desirability bias. In short, text-based models "fail" primarily where the linguistic signal is intrinsically weak or misleading, rather than due to incidental algorithmic shortcomings.

This insight also points to possible ways of improving predictions for these traits: integrating non-linguistic information. For example, in

capturing Conscientiousness, it may be useful to consider behavioural metrics such as the regularity of posting patterns (a conscientious user may post more consistently and thoughtfully) or response punctuality. For Agreeableness, indicators of social interaction such as network structures (a highly agreeable person may exchange more supportive or affectionate messages) may provide additional evidence. In other words, a multimodal or multi-source approach could help compensate for the limitations of text alone (Alsini et al., 2024). However, such approaches introduce new challenges, both technical (integrating heterogeneous sources) and ethical (managing larger volumes of personal data responsibly). These issues will be revisited in the sections on ethical implications.

The key point is that the observed limitations should not be interpreted simply as "errors" to be improved, but rather as reflections of what language can and cannot reveal. Models perform well where the linguistic signal is rich (expressive traits) and struggle where it is weak (internal traits).

3. From Individual Behaviours to Collective Structures

Text-based personality prediction models generate trait estimates that can be used as latent variables, enabling researchers to link individual psychological characteristics with observable behaviours in digital communities. These variables provide a valuable lens through which to analyse the progression from individual behaviour to the formation of homogeneous clusters and, ultimately, to polarisation. In particular, personality differences may influence three interconnected stages:

1. **Network and source selection (selective exposure/avoidance).** Individuals tend to shape their networks and information diets by privileging sources and contacts perceived as compatible with their own beliefs and cognitive style. Those with high Openness to Experience or intellectual curiosity are more likely to incorporate heterogeneous sources and diverse viewpoints into their networks. Conversely, individuals with

stronger cognitive closure or conformity tendencies may avoid content and people that challenge their pre-existing beliefs (*selective avoidance*). These initial choices shape the composition of networks and the degree of exposure to divergent perspectives (Y. Kim et al., 2013; Stroud, 2010a).

2. **Content response (selective engagement).** Once exposed to content, personality traits influence both the likelihood and the nature of engagement (like, comment, share). For instance, high *Neuroticism* may increase sensitivity to emotionally negative or anxiety-inducing content, whereas high *Extraversion* may lead to more frequent interaction with social or celebratory content. These individual responses produce aggregate effects: platform algorithms tend to prioritise and amplify content that generates greater engagement, thereby amplifying material that resonates with the dominant traits of the community (Rathje et al., 2021; Törnberg et al., 2021).

3. **Structural organisation of communities.** The selected connections and generated interactions contribute to the emergence of clusters of users who share similar psychological traits, interests, and worldviews. Within these clusters, internal homogeneity and limited interaction with external groups strengthen *echo chamber* dynamics and increase the risk of both ideological and affective polarisation (Del Vicario et al., 2016; Rabb et al., 2023).

This framework illustrates how personality prediction from language is not merely an individual-level descriptive tool but also an analytical instrument capable of illuminating the mechanisms through which preferences, interactions, and algorithms jointly shape the structure and dynamics of digital communities.

Social network analysis allows these tendencies to be quantified. For example, **modularity** (see Chapter 2, Section 2.1) can be used to identify communities that are highly interconnected internally but weakly connected externally, while **assortativity** (Newman, 2002, 2003) can be calculated for personality traits and ideological orientation. Newman's assortativity coefficient estimates the extent to which

nodes in a network preferentially connect with others sharing the same attribute value (e.g. high or low levels of a given personality trait). The coefficient r ranges from –1 (connections predominantly with dissimilar others) to +1 (connections almost exclusively with similar others), with 0 indicating no correlation. *Positive values* therefore reflect strong *homophily*—the tendency to form ties with similar others—which can reinforce *echo chambers*, while *negative values* indicate greater openness to contacts with dissimilar individuals. Integrating network metrics such as assortativity with psychological profiles makes it possible to assess the extent to which cluster composition is driven by dispositional affinities, and how these affinities may amplify or mitigate ideological and affective polarisation.

Whereas metrics such as modularity and assortativity provide a static snapshot of community structure, personality traits enable a dynamic understanding of the processes that shape these structures, clarifying how individual interactions evolve over time into network configurations that are more or less closed.

3.1 Personality Traits and Echo Chamber Dynamics

Beyond describing the static composition of digital communities, personality traits inferred from text provide valuable insights into the mechanisms that drive their evolution. Certain dispositions act as predisposing factors either towards the formation or the overcoming of echo chambers. Individuals with **high Openness to Experience** and **high Extraversion** tend to form *bridging ties* that connect otherwise separate clusters, thereby increasing informational diversity. By contrast, profiles characterised by **low Openness** or **high Neuroticism** are more likely to form *bonding ties* that reinforce homophily (Bakshy et al., 2015b; Sindermann et al., 2020).

The effect of personality traits on network structure is not independent of the technological context in which interactions occur.

Platform *affordances*—that is, the design features and operational rules that determine which content is displayed, how it is ordered, and how users can interact with it—significantly modulate the relationship between individual dispositions and social network configuration. For instance, algorithmically curated feeds with strong personalisation (e.g.

Facebook, TikTok) tend to recommend content similar to that already liked or shared by the user, thereby amplifying the link between psychological profiles and network homophily: individuals predisposed towards closure (low *Openness*, high *Neuroticism*) are even more frequently exposed to similar content and contacts, reinforcing homogeneous clusters. Conversely, on platforms with chronological ordering and greater incidental exposure to heterogeneous sources, the impact of traits on the formation of homophily in networks tends to diminish, as the diversity of system-suggested content increases the likelihood of connections with dissimilar individuals (González-Bailón & Lelkes, 2023; Guess et al., 2023).

Analysing combinations of traits further highlights profiles at risk of cognitive closure. For example, the combination of low *Openness* and high *Neuroticism* is associated with stronger engagement with confirmatory content and weaker interaction with the out-group (Rathje et al., 2021). These effects are not static: prolonged immersion in ideologically homogeneous clusters can gradually reshape users' language and, in turn, the personality estimates derived from it, in a socialisation process that further reduces internal informational diversity (Bail et al., 2018).

Integrating these perspectives makes it possible to move from a descriptive to a dynamic analysis of *echo chambers*, in which traits, technological affordances, and network structures interact over time, mutually influencing the evolution of polarisation.

3.2 Additional Psychological Mechanisms in Collective Behaviour: Exposure, Conflict, and Cooperation

Beyond shaping the structure of digital social networks and the configuration of homophily in ties, personality traits also influence other fundamental aspects of collective behaviour. In particular, they can modulate cognitive and relational processes that precede, accompany, or counteract the development of polarisation. Among these, three interconnected yet distinct mechanisms can be identified: *selective exposure to information, conflict dynamics,* and *cooperative behaviours* among users. Each represents a critical junction linking individual dispositions, *platform affordances*, and aggregated social outcomes.

Selective exposure refers to the tendency of individuals to seek out, prioritise, or pay more attention to information that aligns with their own belief system, while avoiding sources perceived as dissonant, threatening, or destabilising (Festiger, 1957; Stroud, 2010b). This cognitive mechanism, well documented in the literature on cognitive dissonance, is equally evident in digital contexts, where the personalisation of feeds and sources makes it especially salient.

From a psychological perspective, selective exposure is modulated by specific personality traits. Individuals with *low Openness to Experience* show less inclination to explore novel or divergent content, whereas *high Neuroticism* increases discomfort when confronted with anxiety-inducing information. By contrast, individuals *high in Conscientiousness* tend to build consistent and stable information diets, though these may also become systematically selective.

Computationally, this phenomenon can be modelled by representing the probability of interaction between user and content as a function of *semantic or ideological distance*[27]. In recommender systems, selective exposure may emerge as an unintended by-product of filtering algorithms[28], but can also be mitigated through controlled *diversification strategies*[29] (Kunaver & Požrl, 2017).

The second mechanism, **conflict dynamics**, concerns how individuals react to content or interlocutors perceived as adversarial. Digital platforms, particularly in polarised environments, often incentivise conflictual behaviour: *algorithmic amplification of emotionally charged content*

---

[27] Ideological or semantic distance: a measure of dissimilarity between a user and a piece of content (or between users) within a latent vector space, typically computed using cosine similarity over embeddings generated by models such as Word2Vec, BERT, or LDA topic models (Kulshrestha et al., 2017; Ribeiro et al., 2020).

[28] Recommender filters: models that personalise the content shown to users on the basis of their past behaviour (e.g. collaborative filtering, matrix factorisation, neural recommenders).

[29] Controlled diversification: a recommender strategy that introduces variation into suggested content while preserving relevance to user preferences, typically through redundancy penalisation, multi-objective optimisation, or re-ranking, in order to mitigate selective exposure and echo chambers (Nguyen et al., 2014; Kunaver & Požrl, 2017).

makes messages that provoke outrage, contempt, or ridicule more visible, favouring the spread of flame wars, trolling, and personal attacks (Wahlström et al., 2021).

At the trait level, *high Neuroticism* is associated with greater emotional reactivity, instability, and a tendency to interpret messages as threatening. *Low Agreeableness* correlates with reduced empathic listening and greater propensity for confrontational exchanges. In addition, the combination of *high Extraversion* and *low Conscientiousness* can foster impulsive, attention-seeking communication often aimed at dramatising conflict.

Conflict dynamics can be detected computationally through **stance detection**, **discursive tone classification**[30], and analysis of **semantic escalation** in reply chains. Indicators such as increasing semantic divergence, escalating emotional intensity within threads, or recurrent toxic keywords may serve as predictive signals of escalation. Integrating these metrics with estimated psychological profiles enables the construction of predictive models of relational conflict.

Alongside conflict, some digital environments demonstrate the potential to foster **cooperative and prosocial behaviours**. These include forms of mutual aid, collaborative moderation, spontaneous fact-checking, and emotional support. Cooperation may arise in online groups organised around shared goals, but also in larger networks where users informally assume roles of support or regulation.

Personality traits that facilitate cooperation include *high Agreeableness*, which promotes empathy, willingness to listen, and readiness to help; *high Conscientiousness*, which supports adherence to rules and sustained commitment; and *high Extraversion*, which encourages active participation and the dissemination of positive content. The combination of *Agreeableness and Conscientiousness* appears particularly important for the emergence of prosocial leadership or mediation roles.

---

[30] Stance detection and discourse tone: supervised NLP techniques that classify the viewpoint expressed on a topic (stance: favourable, neutral, opposing) and/or the emotional–rhetorical tone (e.g. moderate, aggressive, ironic), typically using models such as LSTMs, BERT, or fine-tuned Transformers (Hardalov et al., 2022; Mohammad et al., 2016).

Computationally, cooperation can be detected by analysing *patterns of constructive interaction*: frequency of supportive comments, reciprocal appreciation, informative replies, or acts of moderation. Networks may be assessed in terms of **prosocial centrality**[31], distinguishing stabilising users from those who amplify conflict. Models for *detecting empathic or motivational language*[32] further enrich the behavioural map (Van den Bos et al., 2018; Shetty et al., 2024).

Table 6. Psychological mechanisms and personality traits in digital collective behaviour

Mechanism	**Relevant personality traits**	**Computational modelling strategies**
Selective exposure	Low Openness, High Neuroticism, High Conscientiousness	Recommender filters, ideological distance, diversification
Conflict	High Neuroticism, Low Agreeableness, Low Conscientiousness	Stance detection, discursive tone, semantic escalation
Cooperation	High Agreeableness, High Conscientiousness, High Extraversion	Prosocial networks, support patterns, language analysis

The table 6 summarises three behavioural mechanisms—selective exposure, conflict, and cooperation—each associated with specific personality traits and computational strategies for modelling them in digital contexts. Including these psychological mechanisms enables a richer and more nuanced account of online collective behaviours. While network structure describes the architecture of connections and personality traits explain community composition, the mechanisms outlined here clarify how individuals select content, respond to conflict, and engage in cooperation. Integrating these dimensions is a

[31] Prosocial networks and cooperative centrality: network analysis approaches that identify patterns of positive interaction (e.g. support, encouragement, fact-checking) by applying centrality measures (e.g. betweenness, PageRank) to subsets of ties with constructive valence (Van den Bos et al., 2018).

[32] Empathic and motivational language analysis: the classification of textual content based on the presence of prosocial or motivational lexicon, using supervised language models (e.g. LIWC, BERT fine-tuned on help/support corpora) (Shetty et al., 2024).

crucial step towards building predictive and interpretive models that account for the cognitive, emotional, and social complexity of digital communities.

3.3 Joint Prediction of Personality and Polarisation

Analysing echo chambers and personality traits separately provides a rich yet partial picture. The next step is to integrate these dimensions within a single analytical framework, enabling researchers to capture interdependencies and jointly estimate how traits may influence—and at the same time be influenced by—polarisation processes. Studying personality and polarisation as independent phenomena yields useful results, but in many contexts of digital community analysis it is advantageous to model the two dimensions together. The literature suggests, for instance, that certain personality traits (e.g. high Neuroticism or low Openness) can predict a tendency towards polarisation and, conversely, that polarisation can shape personality expression, thereby creating a bidirectional relationship between the two phenomena.

From a statistical perspective, this interdependence can be addressed using:

- **Bivariate models** (e.g. bivariate logistic regression[33]„ Seemingly Unrelated Regression—SUR[34]), which allow for the simultaneous estimation of two dependent variables, exploiting correlation in the residuals to improve efficiency.

[33] Bivariate logistic regression: a statistical model in which two dichotomous dependent variables are estimated simultaneously, accounting for their potential correlation. It is a special case of multivariate binary logistic regression, where each dependent variable follows a Bernoulli distribution but the two regression equations are estimated jointly rather than separately (Agresti, 2013).

[34] Seemingly Unrelated Regression (SUR): a multivariate regression technique introduced by Arnold Zellner (1962)for estimating multiple linear regression equations simultaneously. Each equation has its own dependent variable and set of regressors (which may differ or partially overlap), while the error terms of the different equations are correlated within the same individual/observation.

- **Multivariate mixed models (multilevel models),** which make it possible to incorporate hierarchical levels (e.g. user, thread, community) to account for contextual effects.
- **Multi-output machine learning approaches** (e.g. multi-target Random Forests[35], neural networks with multiple output heads[36]), which can learn shared representations useful for both prediction tasks.

At the semantic level[37], polarisation can also be measured through **embedding distances** (cosine similarity) between groups, by computing the average separation between the vector centroids of posts/comments from different ideological clusters. Combined with predicted personality traits, this approach can reveal patterns such as:

- groups with similar personality profiles but ideologically distant (*thematic polarisation*);
- groups with divergent traits but strong ideological alignment (*selective polarisation*).

The validation of these models requires a testing design capable of assessing both predictive performance for each variable (using standard metrics such as RMSE or F1-score) and the ability to capture the relationship between them (e.g. canonical correlation between predicted and observed values, or correlation coefficients among residuals).

This integrated framework not only improves predictive accuracy but also enables researchers to address deeper quantitative questions

---

[35] Multi-target Random Forest: a generalisation of the classical Random Forest, which predicts a single dependent variable. In the multi-target (or multi-output) setting, each tree is trained to predict multiple dependent variables simultaneously, with each leaf containing a prediction vector rather than a single value. This approach is useful when targets are correlated (e.g. multiple personality traits, various indicators of polarisation) (Breiman, 2001).

[36] Output head: in a neural network, the final layer that produces predictions for a given task. A multi-head neural network includes multiple output layers, each dedicated to a specific target or group of targets (Zhang & Yang, 2021).

[37] Viewing polarisation from the semantic perspective means analysing it not in terms of network structure (who is connected to whom), but through the content and linguistic meaning of the texts produced by users.

such as: *"What proportion of the variance in polarisation is explained by personality traits?"* or *"How does network structure change if the effect of a specific trait is neutralised?"* Such analyses provide a bridge between psychometric modelling and structural analysis of communities, offering concrete tools for monitoring and preventing radicalisation or ideological closure.

In sum, the integration of psychometric analysis, network metrics, and joint predictive approaches moves beyond mere description of individual behaviours, delivering a systemic perspective on the dynamics that shape digital communities and opening the way for targeted interventions to counter polarisation and foster a more open and inclusive information ecosystem.

These findings are not only of theoretical value but also carry practical implications for intervention design. Integrating psychometric profiles with network metrics can support the development of more transparent algorithmic moderation tools, capable of mitigating echo chamber formation and reducing polarisation. At the same time, digital platforms and policy-makers could adopt evidence-based policies to encourage exposure to heterogeneous content, promote bridging ties across communities, and preserve a more open and inclusive information environment.

REFERENCES

Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons Inc.

Al-Fallooji, A. S., & Al-Azawei, A. (2022). Predicting users' personality on social media: A comparative study of different machine learning techniques. *Karbala International Journal of Modern Science*, *8*(4), 617–630.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*(1), i.

Al-Shehri, D. Z., & Al-Qahtani, P. F. M. (2025). Multi-Modal Machine Learning For Comprehensive Public Opinion Analysis In Diverse Regions. *Frontiers in Emerging Artificial Intelligence and Machine Learning*, *2*(06), Article 06.

Alsini, R., Naz, A., Khan, H. U., Bukhari, A., Daud, A., & Ramzan, M. (2024). Using deep learning and word embeddings for predicting human agreeableness behaviour. *Scientific Reports*, *14*(1), 29875.

Alzahrani, H., Acharya, S., Duverger, P., & Nguyen, N. P. (2021). Contextual polarity and influence mining in online social networks. *Computational Social Networks*, *8*(1), 21. https://doi.org/10.1186/s40649-021-00101-3

Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics* (No. arXiv:2008.09470). arXiv. https://doi.org/10.48550/arXiv.2008.09470

Anthis, J. R., Liu, R., Richardson, S. M., Kozlowski, A. C., Koch, B., Evans, J., Brynjolfsson, E., & Bernstein, M. (2025). *LLM Social Simulations Are a Promising Research Method* (No. arXiv:2504.02234). arXiv. https://doi.org/10.48550/arXiv.2504.02234

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarisation. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115

Bakshy, E., Messing, S., & Adamic, L. A. (2015a). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Bakshy, E., Messing, S., & Adamic, L. A. (2015b). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Bama, S., Hema, M. S., Esakkirajan, S., & Nageswara Guptha, M. (2025). A Hierarchical Transformer Network With Label Attention For Personality Prediction By MBTI Classification. *Applied Soft Computing*, 113267.

Barbosa, S., & Milan, S. (2019). Do not harm in private chat apps: Ethical issues for research on and with WhatsApp. *Westminster Papers in Communication and Culture*, *14*(1), 49–65.

Barlow, R. E. (1972). *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. J. Wiley.

Bastos, M. T. (2025). *So Long Twitter, and Thanks for All the Tweets* (SSRN Scholarly Paper No. 5206365). Social Science Research Network. https://doi.org/10.2139/ssrn.5206365

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Bhangale, N. (2025). Personality Prediction Using Machine Learning and Social Media Data: A Myers-Briggs Approach. *International Journal for Research in Applied Science and Engineering Technology*, *13*(4), 4836–4843. https://doi.org/10.22214/ijraset.2025.69336

Blank, G., & Lutz, C. (2017). Representativeness of Social Media in Great Britain: Investigating Facebook, LinkedIn, Twitter, Pinterest, Google+, and Instagram. *American Behavioural Scientist*, *61*(7), 741–756. https://doi.org/10.1177/0002764217717559

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, *3*(null), 993–1022.

Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M. (2025a). Uncovering Digital Trace Data Biases: Tracking Undercoverage in Web Tracking Data.

Communication Methods and Measures, *19*(2), 157–177. https://doi.org/10.1080/19312458.2024.2393165

Bosch, O. J., Sturgis, P., Kuha, J., & Revilla, M. (2025b). Uncovering Digital Trace Data Biases: Tracking Undercoverage in Web Tracking Data. *Communication Methods and Measures*, *19*(2), 157–177. https://doi.org/10.1080/19312458.2024.2393165

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bruns, A. (2021). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Disinformation and Data Lockdown on Social Platforms*, 14–36.

Burnat, F. A. D., & Davidson, B. I. (2025). *The Accountability Paradox: How Platform API Restrictions Undermine AI Transparency Mandates* (No. arXiv:2505.11577). arXiv. https://doi.org/10.48550/arXiv.2505.11577

Cabilio, P., & Masaro, J. (1996). A simple test of symmetry about an unknown median. *Canadian Journal of Statistics*, *24*(3), 349–361. https://doi.org/10.2307/3315744

Castells, M. (2000). *The Rise of The Network Society: The Information Age: Economy, Society and Culture*. Wiley.

Castells, M. (2002). *The Internet Galaxy: Reflections on the Internet, Business, and Society*. OUP Oxford.

Celli, F., & Lepri, B. (2018). Is big five better than MBTI? A personality computing challenge using Twitter data. *Computational Linguistics CLiC-It*, *2018*, 93.

Choong, E. J., & Varathan, K. D. (2021). Predicting judging-perceiving of Myers-Briggs Type Indicator (MBTI) in online social forum. *PeerJ*, *9*, e11382.

Chuang, Y.-S., Goyal, A., Harlalka, N., Suresh, S., Hawkins, R., Yang, S., Shah, D., Hu, J., & Rogers, T. T. (2024). *Simulating Opinion Dynamics with Networks of LLM-based Agents* (No. arXiv:2311.09618). arXiv. https://doi.org/10.48550/arXiv.2311.09618

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, *118*(9), e2023301118. https://doi.org/10.1073/pnas.2023301118

Dai, J., Yan, H., Sun, T., Liu, P., & Qiu, X. (2021). *Does syntax matter? A strong baseline for Aspect-based Sentiment Analysis with RoBERTa* (No. arXiv:2104.04986). arXiv. https://doi.org/10.48550/arXiv.2104.04986

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%253C391::AID-ASI1%253E3.0.CO;2-9

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo chambers: Emotional contagion and group polarisation on facebook. *Scientific Reports*, *6*(1), 37825.

Department of Health, Education, and Welfare. (1979). *The Belmont Report*. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453.

DiMaggio, P., Evans, J., & Bryson, B. (1996). Have American's Social Attitudes Become More Polarized? *American Journal of Sociology*, *102*(3), 690–755. https://doi.org/10.1086/230995

Donkers, T., & Ziegler, J. (2025). *Human-Agent Interaction in Synthetic Social Networks: A Framework for Studying Online Polarisation* (No. arXiv:2502.01340). arXiv. https://doi.org/10.48550/arXiv.2502.01340

Druckman, J. N., Klar, S., Krupnikov, Y., Levendusky, M., & Ryan, J. B. (2021). Affective polarisation, local contexts and public opinion in America. *Nature Human Behaviour*, *5*(1), 28–38. https://doi.org/10.1038/s41562-020-01012-5

Durrheim, K., & Quayle, M. (2025). Human murmuration: Group polarisation as compression in interaction-language dynamics captured by large language models. *European Review of Social Psychology*, 1–40. https://doi.org/10.1080/10463283.2025.2499332

Esteban, J., & Ray, D. (1994). On the Measurement of Polarisation. *Econometrica*, *62*(4), 819–851.

Ferrara, E. (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. *First Monday*. https://doi.org/10.5210/fm.v28i11.13346

Festiger, L. (1957). A theory of cognitive dissonance. *Row, Peterson, New York*.

Fortunato, S., & Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, *104*(1), 36–41. https://doi.org/10.1073/pnas.0605965104

Franzke, A. S., Bechmann, A., Ess, C. M., & Zimmer, M. (2020). Internet Research: Ethical Guidelines 3.0. In *Internet Research* (Vol. 3) [Report]. AoIR (The International Association of Internet Researchers). https://aoir.org/reports/ethics3.pdf

Freelon, D. (2018). Computational Research in the Post-API Age: Political Communication: Vol 35 , No 4—Get Access. *Political Communication*, *35*, 665–668. https://doi.org/10.1080/10584609.2018.1477506

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, *40*(1), 35–41. https://doi.org/10.2307/3033543

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Gayo-Avello, D. (2013). A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. *Social Science Computer Review*, *31*(6), 649–679. https://doi.org/10.1177/0894439313493979

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Golbeck, J., Robles, C., & Turner, K. (2011). Predicting personality with social media. *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, 253–262. https://doi.org/10.1145/1979742.1979614

Goldberg. (1990). An alternative" description of personality": The big-five factor structure. *J Pers Soc Psychol*, *59*(6), 1216–1229.

González-Bailón, S. (2017). *Decoding the social world: Data science and the unintended consequences of communication*. MIT Press.

González-Bailón, S., & Lelkes, Y. (2023). Do social media undermine social cohesion? A critical review. *Social Issues and Policy Review*, *17*(1), 155–180. https://doi.org/10.1111/sipr.12091

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (No. arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Guess, A. M., Malhotra, N., Pan, J., Barberá, P., Allcott, H., Brown, T., Crespo-Tenorio, A., Dimmery, D., Freelon, D., Gentzkow, M., González-Bailón, S., Kennedy, E., Kim, Y. M., Lazer, D., Moehler, D., Nyhan, B., Rivera, C. V., Settle, J., Thomas, D. R., … Tucker, J. A. (2023). How do social media feed algorithms affect attitudes and behaviour in an election campaign? *Science*, *381*(6656), 398–404. https://doi.org/10.1126/science.abp9364

Hardalov, M., Arora, A., Nakov, P., & Augenstein, I. (2022). *A Survey on Stance Detection for Mis- and Disinformation Identification* (No. arXiv:2103.00242). arXiv. https://doi.org/10.48550/arXiv.2103.00242

Harris, J., Germain, J., McCoy, E., & Schofield, R. (2024). Ethical guidance for conducting health research with online communities: A scoping

review of existing guidance. *PLOS ONE*, *19*(5), e0302924. https://doi.org/10.1371/journal.pone.0302924

Hartigan, J. A., & Hartigan, P. M. (1985). The Dip Test of Unimodality. *The Annals of Statistics*. https://doi.org/10.1214/aos/1176346577

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction : with 200 full-color illustrations*. Springer. http://catdir.loc.gov/catdir/enhancements/fy0813/2001031433-t.html

hiQ Labs, Inc. vs. LinkedIn Corp., Sentence No. 17-16783 (Federal Courts 2019). https://law.justia.com/cases/federal/appellate-courts/ca9/17-16783/17-16783-2019-09-09.html

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using bert. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 187–196. https://aclanthology.org/W19-6120/

Hofmann, T. (2013). *Probabilistic Latent Semantic Analysis* (No. arXiv:1301.6705). arXiv. https://doi.org/10.48550/arXiv.1301.6705

Horvát, E.-Á., & Hargittai, E. (2021). Birds of a Feather Flock Together Online: Digital Inequality in Social Media Repertoires. *Social Media + Society*, *7*(4), 20563051211052897. https://doi.org/10.1177/20563051211052897

Hotelling, H., & Solomons, L. M. (1932). The limits of a measure of skewness. *The Annals of Mathematical Statistics*, *3*(2), 141–142.

Howes, R. J., & Carskadon, T. G. (1979). Howes: Test-retest reliabilities of the Myers-Briggs... - Google Scholar. *Research in Psychological Type*, *2*(1), 67–72.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 168–177. https://doi.org/10.1145/1014052.1014073

Hu, N., Pavlou, P. A., & Zhang, J. (Jennifer). (2007). *Why Do Online Product Reviews Have a J-Shaped Distribution? Overcoming Biases in Online Word-of-Mouth Communication* (SSRN Scholarly Paper No. 2380298). Social Science Research Network. https://doi.org/10.2139/ssrn.2380298

Huber, P. J., & Ronchetti, E. M. (1981). Robust statistics, ser. *Wiley Ser Probab Math Stat New York, NY, USA Wiley-IEEE*, *52*, 54.

Humphreys, A., & Wang, R. J.-H. (2018). Automated text analysis for consumer research. *Journal of Consumer Research*, *44*(6), 1274–1306.

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), Article 1. https://doi.org/10.1609/icwsm.v8i1.14550

Interian, R., Marzo, R. G., Mendoza, I., & Ribeiro, C. C. (2023). Network polarisation, filter bubbles, and echo chambers: An annotated review of measures and reduction methods. *International Transactions in Operational Research*, *30*(6), 3122–3158. https://doi.org/10.1111/itor.13224

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (2019). The Origins and Consequences of Affective Polarisation in the United States. *Annual Review of Political Science*, *22*(Volume 22, 2019), 129–146. https://doi.org/10.1146/annurev-polisci-051117-073034

Jelodar, H., Wang, Y., Yuan, C. et al. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*, 15169–15211.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Jiang, K., & Xu, Q. (2023). Analyzing the dynamics of social media texts using coherency network analysis: A case study of the tweets with the co-hashtags of #BlackLivesMatter and #StopAsianHate. *Frontiers in Research Metrics and Analytics*, *8*. https://doi.org/10.3389/frma.2023.1239726

Jockers, M. L. (2015). *R package citation: Syuzhet*. https://github.com/mjockers/syuzhet

Jungherr, A., Rivero, G., Rodríguez, G. R., & Gayo-Avello, D. (2020). *Retooling politics: How digital media are shaping democracy*. Cambridge University Press.

Kim, L. (2023). The Echo chamber-driven Polarisation on Social Media. *Journal of Student Research*, *12*(4). https://doi.org/10.47611/jsr.v12i4.2274

Kim, Y., Hsu, S.-H., & de Zúñiga, H. G. (2013). Influence of social media use on discussion network heterogeneity and civic engagement: The moderating role of personality traits. *Journal of Communication*, *63*(3), 498–516.

Kindsmüller, M. C., & Milz, J. G. (2010). Online Communities and OC Builiding–a Review of Definition and Best Practices. *In: Proceedings of the IADIS International Conferences Collaborative Technologies; Freiburg, Germany, July 26-31.*, 37–40. http://www.imis.uni-luebeck.de/publikationen/OCandOCB-Kindsmueller-Milz-2010-web.pdf

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33–50.

Kopacheva, E., & Yantseva, V. (2022). Users' polarisation in dynamic discussion networks: The case of refugee crisis in Sweden. *PLOS ONE*, *17*(2), e0262992. https://doi.org/10.1371/journal.pone.0262992

Koudenburg, N., & Kashima, Y. (2022). A Polarized Discourse: Effects of Opinion Differentiation and Structural Differentiation on Communication. *Personality and Social Psychology Bulletin*, *48*(7), 1068–1086. https://doi.org/10.1177/01461672211030816

Koutsoumpis, A., Oostrom, J. K., Holtrop, D., Van Breda, W., Ghassemi, S., & de Vries, R. E. (2022). The kernel of truth in text-based personality assessment: A meta-analysis of the relations between the Big Five and the Linguistic Inquiry and Word Count (LIWC). *Psychological Bulletin*, *148*(11–12), 843.

Kulshrestha, J., Eslami, M., Messias, J., Zafar, M. B., Ghosh, S., Gummadi, K. P., & Karahalios, K. (2017). Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 417–432. https://doi.org/10.1145/2998181.2998321

Kunaver, M., & Požrl, T. (2017). Diversity in recommender systems–A survey. *Knowledge-Based Systems*, *123*, 154–162.

Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., & Radford, J. (2021). Meaningful measures of human

society in the twenty-first century. *Nature*, *595*(7866), 189–196. https://doi.org/10.1038/s41586-021-03660-7

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Le Mens, G., & Gallego, A. (2025). Positioning political texts with large language models by asking and averaging. *Political Analysis*, *33*(3), 274–282.

Lee, D. D., & Seung, S. H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (No. arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Lupton, D. (2020). *Data selves: More-than-human perspectives*. Polity Press Cambridge.

Luscombe, A., Dick, K., & Walby, K. (2022). Algorithmic thinking in the public interest: Navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity*, *56*(3), 1023–1044. https://doi.org/10.1007/s11135-021-01164-0

Majumder, N., Poria, S., Gelbukh, A., & Cambria, E. (2017). Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems*, *32*(2), 74–79.

Markham, A., & Buchanan, E. (2012). *Ethical Decision-Making and Internet Research Recommendations from the AoIR Ethics Working Committee (Version 2.0)*. https://www.aoir.org/reports/ethics2.pdf

Martinková, P., Bartoš, F., & Brabec, M. (2023). Assessing Inter-rater Reliability With Heterogeneous Variance Components Models: Flexible Approach Accounting for Contextual Variables. *Journal of Educational and Behavioural Statistics*, *48*(3), 349–383. https://doi.org/10.3102/10769986221150517

Marwick, A. E., & Boyd, D. (2014). Networked privacy: How teenagers negotiate context in social media. *New Media & Society*, *16*(7), 1051–1067. https://doi.org/10.1177/1461444814543995

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52*(1), 81.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(Volume 27, 2001), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Miao, W., Gel, Y. R., & Gastwirth, J. L. (2006). A NEW TEST OF SYMMETRY ABOUT AN UNKNOWN MEDIAN. *Random Walk, Sequential Analysis and Related Topics*, 199–214. https://doi.org/10.1142/9789812772558_0013

Mignemi, G., Calcagnì, A., Spoto, A., & Manolopoulou, I. (2024). Mixture polarisation in inter-rater agreement analysis: A Bayesian nonparametric index. *Statistical Methods & Applications*, *33*(1), 325–355. https://doi.org/10.1007/s10260-023-00741-x

Miller, L. A., & Lovler, R. L. (2018). *Foundations of psychological testing: A practical approach*. Sage publications. https://books.google.com/books?hl=it&lr=&id=bxF7DwAAQBAJ&oi=fnd&pg=PA16&dq=Miller+%26+Lovler,+2018,+Foundations+of+Psychological+Testing&ots=S315Z6c3Eb&sig=ARxSnizu7CTOtF5u3EctpNDshg

Milošević, B., & Obradović, M. (2019). Comparison of efficiencies of some symmetry tests around an unknown centre. *Statistics*, *53*(1), 43–57. https://doi.org/10.1080/02331888.2018.1526938

Mimizuka, K., Brown, M. A., Yang, K.-C., & Lukito, J. (2025). *Post-Post-API Age: Studying Digital Platforms in Scant Data Access Times* (No. arXiv:2505.09877). arXiv. https://doi.org/10.48550/arXiv.2505.09877

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). Semeval-2016 task 6: Detecting stance in tweets. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 31–41. https://aclanthology.org/S16-1003.pdf

Mohammad, S., & Turney, P. (2010). Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In D. Inkpen & C. Strapparava (Eds.), *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of*

Emotion in Text (pp. 26–34). Association for Computational Linguistics. https://aclanthology.org/W10-0204/

Myers, I. B. (2003). *MBTI manual: A guide to the development and use of the Myers-Briggs Type Indicator*. Cpp.

Myers-Briggs Foundation. (2025). *Reliability and Validity of the Myers-Briggs Type Indicator*. https://myersbriggs.org/research-and-library/validity-reliability/

Naz, A., Khan, H. U., Bukhari, A., Alshemaimri, B., Daud, A., & Ramzan, M. (2025). Machine and deep learning for personality traits detection: A comprehensive survey and open research challenges. *Artificial Intelligence Review*, *58*(8), 239. https://doi.org/10.1007/s10462-025-11245-3

NESH. (2021). *Guidelines for Research Ethics in the Social Sciences and the Humanities*. National Committee for Research Ethics in the Social Sciences and the Humanitie (NESH). https://www.forskningsetikk.no/en/guidelines/social-sciences-and-humanities/guidelines-for-research-ethics-in-the-social-sciences-and-the-humanities/

Newman, M. E. J. (2002). Assortative Mixing in Networks. *Physical Review Letters*, *89*(20), 208701. https://doi.org/10.1103/PhysRevLett.89.208701

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E*, *67*(2), 026126. https://doi.org/10.1103/PhysRevE.67.026126

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

Nguyen, D. Q., Vu, T., & Tuan Nguyen, A. (2020). BERTweet: A pre-trained language model for English Tweets. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 9–14). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.2

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: The effect of using recommender systems on content diversity. *Proceedings of the 23rd International Conference on World Wide Web*, 677–686. https://doi.org/10.1145/2566486.2568012

Nielsen, F. Å. (2011). *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs* (No. arXiv:1103.2903). arXiv. https://doi.org/10.48550/arXiv.1103.2903

Pariser, E. (2011). *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, *33*(3), 1065–1076. https://doi.org/10.1214/aoms/1177704472

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York: Bloomsbury Press.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296.

Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., Lu, Z., Zheng, Z., Wang, J. Y., Zhou, D., Gao, C., Xu, F., Zhang, F., Rong, K., Su, J., & Li, Y. (2025). *AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviours and Society* (No. arXiv:2502.08691). arXiv. https://doi.org/10.48550/arXiv.2502.08691

Pittenger, D. J. (2005). Cautionary comments regarding the Myers-Briggs type indicator. *Consulting Psychology Journal: Practice and Research*, *57*(3), 210.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. In P. Nakov & T. Zesch (Eds.), *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 27–35). Association for Computational Linguistics. https://doi.org/10.3115/v1/S14-2004

Rabb, N., Cowen, L., & De Ruiter, J. P. (2023). Investigating the effect of selective exposure, audience fragmentation, and echo-chambers on

polarisation in dynamic media ecosystems. *Applied Network Science*, *8*(1), 78. https://doi.org/10.1007/s41109-023-00601-3

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Ramon, Y., Farrokhnia, R. A., Matz, S. C., & Martens, D. (2021). Explainable AI for psychological profiling from behavioural data: An application to big five personality predictions from financial transaction records. *Information*, *12*(12), 518.

Rathje, S., Van Bavel, J. J., & Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, *118*(26), e2024292118. https://doi.org/10.1073/pnas.2024292118

Rheingold, H. (1993). *The Virtual Community: Homesteading on the Electronic Frontier*. Addison-Wesley. https://doi.org/10.7551/mitpress/7105.001.0001

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. https://doi.org/10.1145/3351095.3372879

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A Model of Text for Experimentation in the Social Sciences. *Journal of the American Statistical Association*, *111*(515), 988–1003. https://doi.org/10.1080/01621459.2016.1141684

Roh, M., & Yang, S.-B. (2021). Exploring extremity and negativity biases in online reviews: Evidence from Yelp. com. *Social Behaviour and Personality: An International Journal*, *49*(11), 1–15. https://doi.org/10.2224/sbp.10825

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, *27*(3), 832–837. https://doi.org/10.1214/aoms/1177728190

Saeteros, D., Gallardo-Pujol, D., & Ortiz-Martínez, D. (2025). Text speaks louder: Insights into personality from natural language processing. *PLoS One*, *20*(6), e0323096.

Safitri, G., & Setiawan, E. B. (2022). Optimization prediction of big five personality in twitter users. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(1), 85–91.

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (No. arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Sarle, W. S. (1983). *Cubic Clustering Criterion*. SAS Institute.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., *et al.* (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One*, *8*(9), e73791.

Shetty, V. A., Durbin, S., Weyrich, M. S., Martínez, A. D., Qian, J., & Chin, D. L. (2024). A scoping review of empathy recognition in text using natural language processing. *Journal of the American Medical Informatics Association*, *31*(3), 762–775.

Shum, K.-M., Ptaszynski, M., & Masui, F. (2025). Big Five Personality Trait Prediction Based on User Comments. *Information*, *16*(5), 418.

Silverman, B. W. (1981). *Using Kernel Density Estimates to Investigate Multimodality. Journal of the Royal Statistical Society, Series B (Methodological), 43, 97-99*. https://www.scirp.org/reference/referencespapers?referenceid=1269060

Sindermann, C., Elhai, J. D., Moshagen, M., & Montag, C. (2020). Age, gender, personality, ideological attitudes and individual differences in a person's news spectrum: How many and who might be prone to "filter bubbles" and "echo chambers" online? *Heliyon*, *6*(1). https://www.cell.com/heliyon/fulltext/S2405-8440(20)30059-1?sf229619943=1

Stewart, R. D., Mõttus, R., Seeboth, A., Soto, C. J., & Johnson, W. (2022). The finer details? The predictability of life outcomes from Big Five domains, facets, and nuances. *Journal of Personality*, *90*(2), 167–182. https://doi.org/10.1111/jopy.12660

Stracqualursi, L., & Agati, P. (2025). Predicting MBTI personality of YouTube users. *Scientific Reports*, *15*(1), 7221.

Stroud, N. J. (2010a). Polarisation and partisan selective exposure. *Journal of Communication*, *60*(3), 556–576.

Stroud, N. J. (2010b). Polarisation and partisan selective exposure. *Journal of Communication*, *60*(3), 556–576.

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 380–385). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1035

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *International Conference on Machine Learning*, 3319–3328. http://proceedings.mlr.press/v70/sundararajan17a.html

Sunstein, C. R. (2001). *Echo Chambers: Bush V. Gore, Impeachment, and Beyond.* Princeton University Press.

Sunstein, C. R. (2002). The Law of Group Polarisation. *Journal of Political Philosophy*, *10*(2), 175–195. https://doi.org/10.1111/1467-9760.00148

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Thapa, S., Shiwakoti, S., Shah, S. B., Adhikari *et al.* (2025). Large language models (LLM) in computational social science: Prospects, current state, and challenges. *Social Network Analysis and Mining, 15*(1), 4. https://doi.org/10.1007/s13278-025-01428-9

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544–2558. https://doi.org/10.1002/asi.21416

Törnberg, P., Andersson, C., Lindgren, K., & Banisch, S. (2021). Modeling the emergence of affective polarisation in the social media society. *Plos One*, *16*(10), e0258259.

Trezza, D. (2023). To scrape or not to scrape, this is dilemma. The post-API scenario and implications on digital research. *Frontiers in Sociology*, *8*. https://doi.org/10.3389/fsoc.2023.1145038

Tromble, R. (2021). Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age. *Social Media + Society*, *7*(1). https://doi.org/10.1177/2056305121988929

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 505–514. https://ojs.aaai.org/index.php/ICWSM/article/view/14517

Van den Bos, W., Crone, E. A., Meuwese, R., & Güroğlu, B. (2018). Social network cohesion in school classes promotes prosocial behaviour. *PloS One*, *13*(4), e0194656.

Vaswani, A., Shazeer, R., Parmar, N., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). *Attention Is All You Need*. Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017), Long Beach, California.

Wahlström, M., Törnberg, A., & Ekbrand, H. (2021). Dynamics of violent and dehumanizing rhetoric in far-right social media. *New Media & Society*, *23*(11), 3290–3311. https://doi.org/10.1177/1461444820952795

Wang, C., Liu, Z., Yang, D., & Chen, X. (2025). *Decoding Echo Chambers: LLM-Powered Simulations Revealing Polarisation in Social Networks* (No. arXiv:2409.19338). arXiv. https://doi.org/10.48550/arXiv.2409.19338

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M. *et al.* (2021). *Ethical and social risks of harm from Language Models* (No. arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Wilson, S. L. (2022). *Social Media as Social Science Data*. Cambridge University Press.

Winsor, C. P. (1932). The Gompertz Curve as a Growth Curve. *Proceedings of the National Academy of Sciences*, *18*(1), 1–8. https://doi.org/10.1073/pnas.18.1.1

Wise, A. F., & Paulus, T. M. (2016). Analyzing learning in online discussions. *The SAGE Handbook of E-Learning Research*, *1*, 270–290.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Q. Liu & D. Schlangen (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Xu, H., Liu, B., Shu, L., & Yu, P. (2019). BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 2324–2335). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1242

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2020). *XLNet: Generalized Autoregressive Pretraining for Language Understanding* (No. arXiv:1906.08237). arXiv. https://doi.org/10.48550/arXiv.1906.08237

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, *57*(298), 348–368. https://doi.org/10.1080/01621459.1962.10480664

Zeng, B., Yang, H., Xu, R., Zhou, W., & Han, X. (2019). Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, *9*(16), 3389.

Zhang, Y., & Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, *34*(12), 5586–5609.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., & Li, X. (2011). Comparing Twitter and Traditional Media Using Topic Models. In P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, & V. Mudoch (Eds.), *Advances in Information Retrieval* (pp. 338–349). Springer. https://doi.org/10.1007/978-3-642-20161-5_34

Finito di stampare nel mese di dicembre 2025
per i tipi di Bologna University Press