

Louvain School of Management

Relationship between filter bubbles and the sharing of fake news on social media: an empirical study on Twitter/X

Authors: Emma Garcia, Margaux Labar
Supervisors: Corentin Vande Kerckhove, Colin Timmers
Academic year 2024-2025
Dissertation for the master of Business Engineering
Master subject and focus: Business Analytics
Daytime schedule

During the preparation of this master's thesis, we utilised AI tools such as ChatGPT for the following purposes:

1. Coding: ChatGPT assisted us in producing and optimising the python scripts used for our data collection and experimental analyses.
2. Writing: ChatGPT helped us with sentence structure and phrasing when writing the thesis.

After using ChatGPT, we diligently reviewed and edited the content produced by the tool. We take full responsibility for the final content presented in this thesis.

We, Emma Garcia and Margaux Labar, affirm that the content of this master's thesis reflects our original work, augmented by the responsible use of AI.

Declared on August 6, 2025.

Abstract:

Over the last few years, the rise of social media has revolutionised the way information is consumed. This transformation offers many advantages, but it also has its drawbacks. This thesis focuses on two major concerns: the sharing of fake news and the enclosure of users in filter bubbles on social media. Drawing on a set of tweets posted by 864 users who shared either fake or real news according to the Gossip Cop website, we investigate four research questions: Are users who share fake news exposed to less diverse content? Do they tend to have stronger opinions? Do they tend to share more recommended content rather than original tweets? Are there any features that distinguish this type of user?

Using a quantitative approach, we combine topic modelling and Rao's diversity measure to assess content diversity, sentiment analysis to evaluate intensity of opinions, and statistical comparisons to analyse user publication behaviour and explicit features. Our results show that users who share fake news are significantly more likely to be trapped in a filter bubble – as they are faced with less thematically diverse content – and to repost recommended content through retweets and quoted tweets, suggesting stronger algorithmic influence. Additionally, these users tend to be more active on the platform in terms of posting, liking, and following, and their posts tend to generate more retweets. However, no evidence suggests that their opinions are more extreme.

These findings contribute to a better understanding of the relationship between the spread of fake news and the filter bubble phenomenon on social networks. They highlight the need to place greater emphasis on diversity in algorithms and performance indicators, and to invest more in preventive detection strategies and user awareness to mitigate the risks posed by these two phenomena.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management

Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique
www.uclouvain.be/lsm

Acknowledgements

First of all, we would like to thank Professor Vande Kerckhove for his guidance during this thesis. His valuable advice helped us to overcome moments of doubt and move forward more serenely in the development of this thesis.

Secondly, we would also like to especially thank Colin Timmers. His availability, regular feedback and knowledge of the subject greatly contributed to our work.

Finally, we would like to thank our respective families for their involvement and the support they provided us during these 5 years of study.

Table of contents

1. Introduction.....	1
2. Filter bubble and fake news: literature review.....	4
2.1. Recommender systems.....	4
2.1.1. Approaches used to generate recommendations in recommender systems	4
2.1.2. Algorithmic dynamics and information diversity on social platforms	7
2.2. Filter bubbles and fake news: definitions, mechanisms and limitations	10
2.2.1. Filter bubbles.....	10
2.2.2. Fake news.....	16
2.3. Link between filter bubbles and fake news: existing studies and motivation for our research.....	20
3. Link between filter bubbles and the sharing of fake news: a practical approach	25
3.1. Methodology	25
3.1.1. Research questions to be explored and approach used	25
3.1.2. Data collection	26
3.1.3. Calculation of thematic diversity scores.....	31
3.1.4. Calculation of sentiment scores.....	37
3.1.5. Method used to analyse user publication behaviour.....	40
3.1.6. Method used to analyse user explicit features	41
3.2. Experimental results.....	41
3.2.1. Descriptive analysis of thematic diversity scores.....	42
3.2.2. Descriptive analysis of sentiment scores	44
3.2.3. Descriptive analysis of user publication behaviour	46
3.2.4. Analysis of user features	47
4. Discussion.....	54
5. Managerial implications	58
6. Conclusion.....	60
7. Appendices.....	61
8. References.....	70

List of tables

Table 1: Comparison of the different recommender system approaches	7
Table 2: Reasons why people believe fake news	19
Table 3: Extract from the “news” table in the database	26
Table 4: Extract from the “collected_tweets_fake” table in the database.....	27
Table 5: Extract from the “users” table in the database	29
Table 6: Extract from the “user_timelines” table in the database	30
Table 7: Examples of tweet-topics assignations.....	35
Table 8: Comparison between the raw score returned by each lexicon for a term and the associated revised score.....	39
Table 9: Summary of the hypotheses, their statistical test and associated p-value	42

List of figures

Figure 1: Consequences and risks of algorithmic personalisation on users	10
Figure 2: Filter bubble loop.....	12
Figure 3: Summary of the data collection phase	26
Figure 4: Summary of Bache et al.'s method applied to our study.....	33
Figure 5: C_v coherence scores based on the numbers of topics.....	35
Figure 6: Distribution of users' Rao diversity scores by label	43
Figure 7: Distribution of users' sentiment scores by label	45
Figure 8: Distribution of users' proportions of retweets/quoted tweets by label.....	47
Figure 9: Log-log cumulative rank curves of the number of tweets posted	50
Figure 10: Log-log cumulative rank curves of the number of tweets liked	51
Figure 11: Box plots of the number of followings.....	52
Figure 12: Comparison of TFF ratio distributions by label	52
Figure 13: Box plots of the number of retweets	53
Figure 14: Log-log cumulative rank curves of the number of retweets	53

List of appendices

Appendix 1: Examples of spam users on Twitter/X	61
Appendix 2: User with a limited number of posts in his/her timeline	62
Appendix 3: Example of a user's Twitter/X timeline.....	63
Appendix 4: Extract from the output of the MALLET command providing the most representative words for each topic.....	63
Appendix 5: Normal quantile-quantile plots of users' Rao diversity scores by label	64
Appendix 6: Normal quantile-quantile plots of users' sentiment scores by label	64
Appendix 7: Normal quantile-quantile plots of users' registration time by label	64
Appendix 8: Normal quantile-quantile plots of users' number of tweets by label.....	65
Appendix 9: Normal quantile-quantile plots of the number of tweets liked per user by label	65
Appendix 10: Normal quantile-quantile plots of users' number of followers by label.....	65
Appendix 11: Normal quantile-quantile plots of users' number of followings by label	66
Appendix 12: Normal quantile-quantile plots of users' tweeting time range by label.....	66
Appendix 13: Normal quantile-quantile plots of the average number of replies generated per user by label	66
Appendix 14: Normal quantile-quantile plots of the average number of retweets generated per user by label	67
Appendix 15: Percentage of verified users by label.....	67
Appendix 16: Box plots of user registration time	67
Appendix 17: Box plots of the number of followers.....	68
Appendix 18: Box plots of the tweeting time range.....	68
Appendix 19: Log-log cumulative rank curves of the tweeting time range	68
Appendix 20: Box plots of the number of replies	69
Appendix 21: Log-log cumulative rank curves of the number of replies	69

1. Introduction

Today, the World Wide Web and social platforms such as Facebook and Twitter (newly named X) have revolutionised the way information is disseminated. Traditional news media such as newspapers and magazines have evolved into digital formats in the form of blogs, social feeds and other online platforms (Ahlers, 2006). This transition has many advantages. Information is accessible in real time. This accessibility facilitates social, economic and political decision-making (Ahmad et al., 2020). Users of the various social media platforms also enjoy the ability to discuss, share their feelings and ideas, and debate various issues relating to democracy, education and health (Ahmad et al., 2020; Hosseinimotlagh & Papalexakis, 2018).

Nevertheless, massive information flows often lead to the propagation of inaccurate or misleading information. Some entities can create biased opinions, manipulate mindsets and spread nonsense for profit (Ahmad et al., 2020). As explained by S. Yang et al. (2019), people are faced with increasingly noisy and inaccurate information on social media due to its easy access. This phenomenon has a specific name, that of fake news, which consists of deliberately erroneous or deceptive information presented in the form of a news, written for various purposes (Allcott & Gentzkow, 2017; Gelfert, 2018; Meel & Vishwakarma, 2019; Zhang & Ghorbani, 2020; Zhou & Zafarani, 2020). This was the case, for example, during the 2016 US elections (Holan, 2016) or during the conflict between Ukraine and Russia (Shin et al., 2023). The fact that, in this day and age, people can no longer distinguish between what is true and what is false poses problems in various sectors. In fact, these problems will have an impact on the political, sporting, scientific and financial fields, with, for example, the collapse of the market due to rumours (Kogan et al., 2018; Lazer et al., 2018). In their work, Shin et al. (2023) observe that many people suffer from news that mixes erroneous information with facts due to a certain lack of knowledge. Despite the fact that some news platforms take steps to disprove these false reports, they have been read by millions of people, damaging the public perception of the subjects concerned.

Against this backdrop of information overload and the difficulty of identifying reliable content from fake news, another concept complicates matters: the filter bubble. This term was first defined by Pariser (2011) as the algorithmic personalisation of online content. Recommendation systems offer users information based on their past behaviour, preferences and interactions. Filter bubbles thus create digital environments in which users are exposed to similar content, sources or viewpoints, often to the detriment of divergent opinions (Nguyen et al., 2014; Sunstein, 2001; Tetlock, 2017). One consequence of these filter bubbles is to reduce

the diversity of information and reinforce people's existing beliefs. This will have the impact of limiting critical thinking and constructive debate (Eady et al., 2019; Schomakers et al., 2020). The interaction between the spread of fake news and filter bubbles is a particularly worrying subject. Indeed, Zimmer et al. (2019) observe that when users are exposed to homogeneous and emotionally engaging content, they become more vulnerable to fake news. Furthermore, recommendation algorithms are designed to maximise engagement. As a result, they tend to favour sensational content that is more likely to go viral. These characteristics are generally associated with fake news (Lazer et al., 2018; Meel & Vishwakarma, 2019; Zhang & Ghorbani, 2020). Thus, filter bubbles passively shape what users see and could actively participate in the amplification and polarisation of misleading narratives. All these consequences create an environment in which false information circulates and public opinion becomes more fragmented (Andrea, 2022; Karlsen et al., 2017; Pariser, 2011; Rowland, 2011; Sunstein, 2001).

While much research has focused on fake news or filter bubbles separately, few studies have analysed the interaction between these two concepts. Yet understanding the relationship between the two phenomena is important for developing effective solutions to misinformation. It is precisely this gap that our research aims to fill. In this thesis, we will seek to determine whether a link can be established between the sharing of fake news on social networks and enclosure in a filter bubble. In addition, we will seek to complement existing research by trying to learn more about the behaviours that characterise users who share fake news online.

Through the study of empirical data that we have collected on Twitter/X, we will try to answer various questions. Firstly, we will look at the following question: "Are users who share fake news on social media more likely to be trapped in a filter bubble?". To do this, we will seek to determine whether users who share fake news are confronted with less thematically diverse content. Secondly, we will also look at whether these users tend to have stronger opinions. Next, we will assess whether users who share fake news tend to share significantly more recommended content than other users, by analysing the proportion of retweets and quoted tweets that appear in their Twitter/X timeline. Finally, we will study the explicit features of users, with the aim of finding out whether there are any attributes that make it possible to distinguish users who share fake news from those who share real news on social networks.

This research is particularly relevant in today's context, where the spread of online misinformation poses a serious threat to democratic debate and public trust in media (Pennycook & Rand, 2019; Vosoughi et al., 2018). The increasing reliance on algorithmic filtering and personalised content amplifies the risk that users remain restricted within

homogeneous environments (Chueca Del Cerro, 2024; Cinelli et al., 2021), reinforcing their beliefs and potentially encouraging the sharing of misleading information (Dahlgren, 2021; Del Vicario et al., 2017). Despite the growing number of studies on the concepts of fake news and filter bubbles examined separately, the link between the two remains largely unexplored in real-world settings. By combining topic modelling, sentiment analysis, and behavioural profiling on real user data, our thesis offers a more holistic and empirical approach to this complex phenomenon. Beyond academic interest, we aim to identify the managerial implications that arise from our findings, discussing platform design, management strategies and media literacy initiatives, with the broader aim of fostering a healthier information ecosystem.

To this end, the thesis follows a clear structure. Section 2 provides a review of the existing literature on the subject of recommender systems, fake news and filter bubbles. This section discusses recommender systems, looking at their approaches and their role in personalising information on social networks. It also defines the concepts of filter bubble and fake news, analyses their mechanisms and their potential effects on the perception and dissemination of information, and takes a critical approach. Finally, the section looks at the link between the two concepts by highlighting existing studies and positioning our research within them.

Section 3 covers the practical approach we adopted to study the link between filter bubbles and the sharing of fake news on social networks. It details the methodology we employed, providing in-depth explanations of the data collection phase and presenting the different methods we implemented to evaluate each research question. It also develops the experimental results we obtained to answer our various research questions.

Section 4 concludes our research with a discussion of the results, also identifying the various limitations of our study and suggesting ideas for future work.

Section 5 highlights the managerial implications of our research.

Finally, the last part of this work provides a conclusion to our thesis. Next come the appendices and the bibliographical references on which our work is based.

2. Filter bubble and fake news: literature review

2.1. Recommender systems

In a world where information sources are multiplying, users are increasingly faced with information overload on networks. This phenomenon occurs when the amount of data available exceeds the user's processing capacity (Carlson, 2007). Faced with an increasing amount of information, users are struggling to make clear and effective decisions. One proposed solution is the use of recommendation systems. These systems are powerful tools for reducing information overload (Kumar & Sharma, 2016). Through the use of this kind of system, users benefit from improved decision making, based on the experience or preferences of other users (Ricci et al., 2011).

The definition of recommender systems has evolved in recent years. Burke (2002), subsequently, defines the concept as a set of tools and techniques providing personalised suggestions to the user on items likely to be of interest to them. This definition no longer involves how the recommendations are generated, but focuses on the personalised effect for the user. Lü et al. (2012) add that the aim of this system is to predict a user's future preferences from input data, often based on past user evaluations. Finally, the model was formalised by Adomavicius & Tuzhilin in 2005. This formalisation makes it possible to include typical database or personalised information retrieval applications, as well as systems broader than those based on social aggregation.

In addition to helping users navigate vast amounts of information, recommender systems also play a central role in shaping the way content is presented on digital platforms. In order to understand how users may become trapped in a filter bubble as well as being more exposed to misleading information, it is essential to explore these algorithmic mechanisms. Understanding how these systems operate is crucial for analysing the potential role they might play in the reinforcement of informational narrowing, the formation of filter bubbles, and the spread of content online.

2.1.1. Approaches used to generate recommendations in recommender systems

As explained above, recommender systems focus on users' past evaluations, so the main task of recommender systems is to predict the rating of an unrated item and recommend those with the best ratings (Melville & Sindhvani, 2011). To personalise content by exploiting user interests, behavioural data and contextual information (Oechslein & Hess, 2013), personalisation systems generally rely on two filtering approaches: content-based filtering and

collaborative filtering. Table 1 shows the differences between these two filtering approaches, as well as a comparison with a hybrid approach combining these two types of filtering.

Content-based filtering is a method that consists of comparing the characteristics of an item's content with those of the user's known interests (Melville et al., 2002). This filtering works independently of the opinions of other users. This approach makes it possible to recommend items similar to those enjoyed by the user (Ricci et al., 2011). It is effective for text documents whose content can be easily analysed. Among the advantages of this filtering method is its ability to provide personalized recommendations tailored to each user's specific preferences and past interactions. However, the method has certain limitations. Given that the approach is based solely on the characteristics of the items, if these do not reflect what the user really likes, then the suggestions may be inappropriate. This is due to the lack of available data. The system cannot accurately distinguish a user's preferences if the available data is not rich or accurate enough (Han et al., 2012; Mohanty et al., 2020). There is also a risk of over-specialisation, as recommendations are based on similar items previously liked by the user, which limits the diversity of suggestions (Aggarwal, 2016; Mohanty et al., 2020).

Collaborative filtering allows items to be recommended to users based on their ratings or usage patterns, without the need for external information about users or items. The main approach to collaborative filtering focuses on neighbourhood methods, i.e. it is based on relationships between users or between items (Schafer et al., 2007). We can find memory-based collaborative filtering, which makes it possible to recommend items appreciated by users similar to the target or similar to items already appreciated by the user (Liu et al., 2016). Recommendations are therefore generated either by identifying users with similar preferences (user-based filtering) or by finding items that are similar to those the user has previously liked, based on shared patterns of appreciation across other users (item-based filtering). There is also model-based collaborative filtering, which uses predictive models trained on the data to discover patterns (Liu et al., 2016). For example, latent factor models are used to represent users and items in a latent factor space. Factors are automatically inferred from user ratings (Koren et al., 2022). This type of collaborative filtering is effective when analysing complex content because it relies on historical information such as user preferences. Actual content is not necessarily required when this method is used (H.-N. Kim et al., 2011).

Collaborative filtering is widely used for different types of content such as movies, books, music, etc. (Resnick et al., 1994). However, collaborative filtering has certain limitations. The

authors Schein et al. (2001, 2002) were able to identify cold-start problems: cold-start items and cold-start users. These cold-start problems exist because it is impossible for the recommendation system to make recommendations. Cold-start items mean that an item cannot be recommended until it has received sufficient ratings from users. The system will therefore focus on the similarity of content between different items in order to make recommendations between similar recommended items. This filtering approach may therefore suffer from over-personalisation (Kumar & Sharma, 2016). Cold-start users happen when the system does not yet know the user's preferences, behaviour or interests. Knowing that the filtering relies on past interactions, the method cannot be applied effectively. This lack of data makes it difficult to personalise recommendations or predictions accurately (Melville & Sindhvani, 2011; Schein et al., 2002). We can therefore add that there is a degree of data sparsity, as there is not enough data for some users or items. It also leads to scalability problems, as the system must handle an increasing number of users, items, and interactions efficiently, which requires substantial computational resources and sophisticated data management techniques. In addition, some users are unreliable because they give false ratings or are spammers (Burke, 2002). This type of user, known as the "grey sheep", is unhelpful if the system wants to make consistent recommendations (Su & Khoshgoftaar, 2009).

In order to exploit the combined advantages of these two filtering approaches, some researchers have proposed combining content-based filtering with collaborative filtering to reduce their disadvantages. Many researchers have proposed their own hybrid approach (Burke, 2002). These include techniques such as mediating user models by integrating data from different systems (Berkovsky et al., 2008) or transforming a collaborative filtering model into a content-based filtering model (Berkovsky et al., 2009). Another technique is to use an approach with Bayesian classification to predict missing ratings (content-based filtering) (Melville et al., 2002). It can also be the recalculation of ratings by collaborative filtering (Melville et al., 2002) or an inverse approach (first collaborative filtering then content-based filtering on the results obtained) (Salter & Antonopoulos, 2006). This hybrid method therefore combines several recommendation techniques in order to obtain a final result. The combination of these two filtering techniques provides better shared knowledge, contributing to better recommendations. As cited by Geetha et al., (2018), the increase in knowledge makes it possible to explore new ways of extending the underlying collaborative filtering algorithms to content data and content-based algorithms to user behaviour data.

Feature	Content-Based Filtering	Collaborative Filtering	Hybrid Filtering
Data Used	Item features + user profile	User-item interactions (ratings, behaviour)	Combines both
Dependency on Other Users	Independent of others	Based on other users' opinions	Combines both
Over-Personalisation Risk	High	Lower (diverse inputs from others)	Mitigated through hybridisation
Model Types	Profile-based (user interest matching)	Memory-based / Model-based	Mixed (e.g., cascade, ensemble, switching, feature-combination)
Limitations	Narrow recommendations, cold start	Sparsity, scalability, unreliable users (gray sheep)	Increased complexity

Table 1: Comparison of the different recommender system approaches

2.1.2. Algorithmic dynamics and information diversity on social platforms

This section adopts a complex systems perspective to explore how recommendation algorithms, feedback loops, and information propagation models influence users' exposure to content. It highlights the mechanisms behind algorithmic personalisation, the types of diversity relevant to democratic discourse, and the potential risks posed by algorithmic filtering – particularly the amplification of echo chambers and filter bubbles.

Social networks are increasingly perceived as complex systems, an idea widely explored in network sciences and systems theory (Barabasi, 2024). These systems are characterised by dynamic interactions between their components – like users, contents, social connections, etc. –, which give rise to emergent phenomena that are difficult to predict (Meadows, 2008).

Social platforms are also influenced by feedback loops, which reinforce or attenuate certain engagement dynamics. Analysing these mechanisms remains complex, notably due to the lack of transparency of algorithms and limitations on access to platform data (Narayanan, 2023). Bakshy et al. (2015) have shown that algorithmic personalisation of news feeds significantly influences users' exposure to certain ideas. This is because the algorithm reacts to user signals. If a user ignores divergent content, the algorithm stops proposing it, which will reinforce existing attitudes (Stroud, 2008). Recommendations will therefore reinforce the user's bias and reduce the diversity of information appearing in their news feed. Messing & Westwood (2014) confirm this point by adding that a user showing an interest in diversity is more likely to have the algorithm favour greater diversity in its recommendations. The risks of the personalisation algorithm can take the form of two well-known theories: filter bubbles and echo chambers. The concept of filter bubble popularised by Pariser (2011) is the result of excessive personalisation

of online content. Users find themselves locked into a bubble of information that reduces exposure to divergent opinions or diverse content. According to Cinelli et al. (2020), an echo chamber is an online social group where opinions on a controversial topic are highly polarised. It is also marked by homophily, meaning members tend to interact mainly with others who share their views. This phenomenon leads to repeated exposure to content that conforms to pre-existing beliefs, reinforcing individual convictions and limiting openness to divergent points of view (Cinelli et al., 2020).

Algorithms play a central role in the way information circulates on these platforms. While content processing algorithms adapt to new trends, information propagation algorithms determine which users are exposed to which types of content. Among these, recommendation algorithms are particularly influential, as they shape information flows based on users' previous behaviours in order to maximise their engagement (Zeng et al., 2021). These systems are not limited to social networks; they are also at the heart of online streaming and commerce platforms, such as Netflix and Amazon (Gomez-Uribe & Hunt, 2016).

Three fundamental models of information propagation are often cited in the scientific literature: the subscription model, the network model and the algorithmic model. The subscription model works in a simple way: a user receives publications directly from accounts to which they have subscribed, which is reminiscent of the way traditional newspaper subscriptions work (Narayanan, 2023).

In the network model, dissemination is based on the propagation of content through connections between users. This encourages the creation of information cascades and has long been the basis of Twitter's operation, particularly before 2016. This model highlights the influence of certain highly connected users in the dissemination of content (Narayanan, 2023).

Finally, the algorithmic model relies on behavioural predictions to adapt the content displayed to each user. Unlike the two previous models, it does not rely directly on subscriptions or interactions between users, but on algorithms that dynamically adjust content according to potential engagement (Narayanan, 2023). This model is present on platforms such as TikTok and YouTube, which rely heavily on adaptive algorithms (Zeng et al., 2021).

In parallel, a growing proportion of consumers depends on personalised digital environments (Haim et al., 2017). The algorithmic personalisation of news depends on explicit personalisation and implicit personalisation. Explicit personalisation is based on the preferences indicated by

the user, whereas implicit personalisation is based on the analysis of browsing behaviour (Thurman & Schifferes, 2012; Zuiderveen Borgesius et al., 2016). These algorithms will automatically select content, which is different from human editors (Haim et al., 2017).

The diversity resulting from what these algorithms propose is essential for the quality of information and a well-informed public (McQuail, 1993; Strömbäck, 2005). There are often three forms of diversity: diversity of sources, diversity of content and diversity of viewpoints. Diversity of sources consists of the variety of media and people quoted (Voakes et al., 1996). It includes the variety of affiliations of the actors quoted, but also encompasses the diversity of media included in the consumption of news (Thurman, 2011). Diversity of content corresponds to the variety of subjects covered and ideally includes all the perspectives available on a given subject (R. M. Entman & Wildman, 1992; Voakes et al., 1996). When we talk about perspectives, we are implying the fact of referring to the interpretative frameworks applied to a subject. More specifically, diversity of viewpoints corresponds to the presence in the media of different interpretations of a subject, and not only to a diversity of sources or tones. Baden & Springer (2015), following the work of Wetherell & Potter (1988), and Entman (1993), define it as the coexistence of non-commensurable interpretative frames, derived from distinct interpretative repertoires. According to Donati (1992) and Fairclough (2003), these repertoires are coherent semantic structures based on logics of action and evaluation, which Boltanski & Thévenot (2006) have systematised into a typology of justifications. Two frameworks will then be considered as diverse if they cannot be reduced to the same logic or the same repertoire. It is important to note that citing different sources, as Bennett (1996) and Tat et al. (2002) point out, does not guarantee real diversity of interpretation since several players may share the same dominant framework. Diversity of viewpoints is the most difficult form of diversity to produce and measure (Baden & Springer, 2017). It makes it possible to ensure a democratic “marketplace of ideas” (Strömbäck, 2005). The capacity of algorithms to support such diversity has been increasingly called into question. Pasquale (2015) questioned the ability of algorithms to guarantee diversity. This gave rise to the filter bubble hypothesis, which suggests that algorithms favour economic commitment over diversity. According to Pariser (2011), they favour the content most likely to interest a given user. One example cited is that a user who consumes a lot of sports news will receive even more sports content to the detriment of other subjects (Haim et al., 2017).

In summary, this section has explained how social media platforms operate as complex systems, where algorithms play an important role in the choice of content seen by internet users. These

algorithms, particularly those that recommend content, learn from user behaviour and tend to show more of what a person already likes. Over time, this can limit the diversity of content the users are exposed to. Instead of discovering different points of view or new subjects, users often see the same type of content. This reduces the diversity of information that is important for staying well-informed and understanding different opinions. This is how online spaces can become filter bubbles or echo chambers. Figure 1 summarises the risks and consequences that can arise from algorithm personalisation. This brings us to the next part: the idea of the filter bubble, which describes how algorithms can trap users in narrow information universes by showing them only what they already agree with.

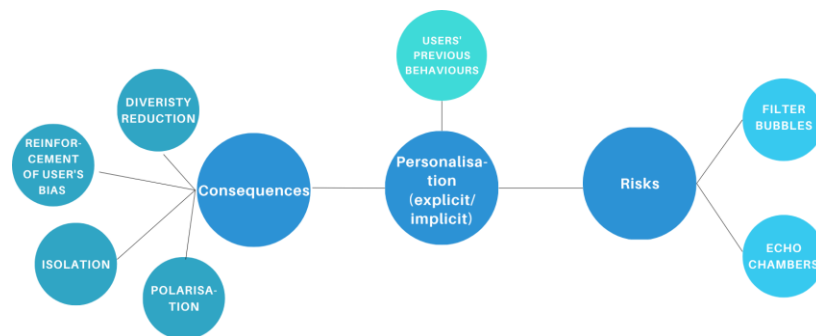


Figure 1: Consequences and risks of algorithmic personalisation on users

2.2. Filter bubbles and fake news: definitions, mechanisms and limitations

2.2.1. Filter bubbles

2.2.1.1. Definition and origin

One potential consequence of recommendation systems is the creation of filter bubbles within social networks. Recommendation systems will tend to reinforce or attenuate them, influencing the informational opening or closing of the Internet (Areeb et al., 2023). But what does this concept mean?

In 2011, Eli Pariser raised growing concerns about the development of the web and the personalisation of its content. He highlights some of the unintended effects of this personalisation, which is becoming ubiquitous and barely visible in the critical evaluation of information. One of the main risks he identifies is the fragmentation of users into isolated informational spaces, a phenomenon he describes as the “filter bubble”. This concept, which he developed, describes a personalised information environment in which each user evolves in a unique way (Pariser, 2011).

Pariser explains that this phenomenon is based on algorithmic personalisation mechanisms that filter and select content according to the user's preferences and past interactions (Pariser, 2011). This personalisation locks individuals into a restricted environment, reducing their creativity

and reinforcing their pre-existing beliefs (Andrea, 2022). Sunstein (2001) reinforces this perspective, arguing that algorithmic personalisation undermines shared experiences and, as a result, weakens the ability of diverse societies to address common challenges. Some authors add that personalisation can reinforce users' cognitive biases by exposing them only to content aligned with their preferences (Nguyen et al., 2014; Tetlock, 2017).

Let's take a closer look at Pariser's thesis to better understand the concept of the filter bubble. He argues that personalisation restricts the diversity of information, particularly through search engines and social networks, which filter content without the user being aware of it (Pariser, 2011). For example, Google customises search results according to the user's profile, Facebook tracks interactions to optimise advertising targeting and Amazon exploits purchase correlations to anticipate preferences. Exploiting private data enables these companies to refine their recommendations and optimise targeted advertising. The more extensive and precise the data collected – including detailed information such as users' likes, browsing times, and inferred preferences – the more valuable it becomes for advertisers. However, users are often unaware of this constant exposure (Samuels, 2012).

Pariser (2011) points out that algorithms build a digital profile based on past interactions, limiting access to other perspectives. Two types of personalisation have been identified.

On the one hand, algorithmic personalisation is based on the analysis of behavioural data such as browsing history, pages consulted, clicks and time spent on certain content. It also incorporates metadata such as the date of publication, the type of document and the number of interactions (likes, shares). The algorithm processes this information to suggest similar content, often without the user noticing (Bozdag, 2013; Helberger et al., 2018).

On the other hand, explicit personalisation or “customisation” is based on voluntary action by the user, who chooses to filter content (Sundar & Marathe, 2010), for example by liking a Facebook page or following an account on Twitter/X. Unlike algorithmic personalisation, this personalisation is consciously initiated by the user (Andrea, 2022).

Different authors have identified several effects and risks associated with filter bubbles. Although personalisation does not systematically lead to a bubble, in extreme cases it can create an environment where users are exclusively exposed to content that reinforces their existing opinions. This mechanism, known as “amplification”, is based on a circular dynamic: the more a user consults a particular type of content, the more the algorithm suggests, reinforcing the algorithmic spiral. This process limits confrontation with divergent points of view and reduces

informational diversity (Andrea, 2022; Bakshy et al., 2015; Flaxman et al., 2016; Pariser, 2011). This dynamic loop process is shown in Figure 2.

However, Andrea (2022) qualifies this view by pointing out that the filter bubbles are not totally hermetic. Content contrary to the user's preferences can enter them, provided it comes from sources deemed 'acceptable' by the algorithm.

Moreover, the societal impact of these filter bubbles is significant. One of the main risks is social fragmentation: users evolve in distinct informational spheres, limiting exchanges between groups with divergent opinions. This compartmentalisation encourages intellectual isolation and reduces the diversity of public debate (Eady et al., 2019; Schomakers et al., 2020). By restricting access to varied information, filter bubbles represent a major challenge for democracy. When combined with echo chambers (content shared in homogeneous circles), they exacerbate public polarisation and fragmentation, making societal dialogue more difficult (Andrea, 2022; Karlsen et al., 2017; Pariser, 2011; Rowland, 2011; Sunstein, 2001).

We can therefore see that filter bubbles have different impacts on users. Their existence tends to reinforce existing beliefs, which polarises information (Nguyen et al., 2014; Tetlock, 2017). There will also be the creation of a group effect within which a radicalisation of opinions will take place (Gharahighehi & Vens, 2023). Popular content is also recommended more, reducing diversity. The authors Fleder & Hosanagar (2009) and Lee & Hosanagar (2014) refer to this consequence as the Matthew effect. Filter bubbles also have an impact on the socio-political situation as they influence political ideas, leading to a weakening of democracy (Eady et al., 2019; Schomakers et al., 2020). Finally, these bubbles could amplify misinformation, encouraging the spread of fake news (Zimmer et al., 2019).

So while the personalisation of content offers advantages in terms of relevance and convenience, it also entails dangers that it is essential to understand and mitigate in order to preserve access to diverse information and informed public debate.

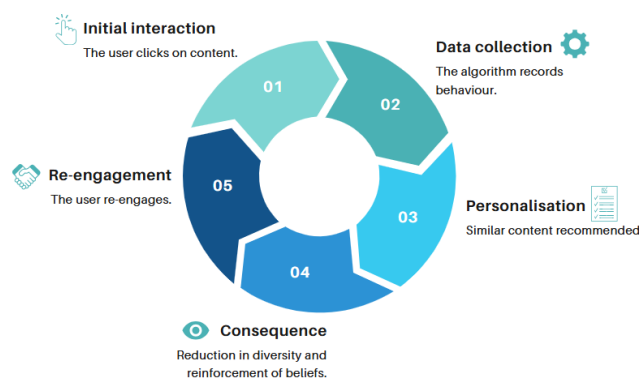


Figure 2: Filter bubble loop

2.2.1.2.Criticisms and limitations of the filter bubble concept

Having defined the concept of filter bubble, it is interesting to examine this phenomenon in greater depth. In this section, we will explore the various criticisms raised by different authors regarding the filter bubble hypothesis. In fact, many articles criticise Pariser's theory, so it is worth looking at whether or not there really is a "bubble" within social networks.

As mentioned several times in the previous sections of this work, users look for information with which they agree and avoid information that contradicts them. Frey (1986) already identified the theory of selective exposure in one of his analyses. Using this theory, he shows that people tend to seek out information that is compatible with their beliefs, but rarely avoid information that contradicts them. We can therefore say that all information consumption is selective by necessity. If we analyse Pariser's concept (2011), we see that he states that personalisation amplifies confirmation bias. This amplification tends to transform the media into a reflection of personal preferences. Many studies confirm a slight preference for information that supports users' ideas (Cotton, 1985; D'Alessio & Allen, 2007; Hart et al., 2009) but, however, there is little evidence of active avoidance of contrary information. Numerous researchers have identified that the motivation to seek information that supports is stronger than the motivation to avoid information that confronts (Fischer et al., 2011; Garrett, 2009; Munson & Resnick, 2010). In their research, Newman et al. (2017) found that 40% of social network users claim to see information that does not interest them, or at least that they are not looking for. Hart, Richardson, et al. (2020) go so far as to assert that the more confident a person is in his beliefs, the more he is exposed to contrary information. We can therefore say that people look for information that supports them but do not do everything possible to avoid information that contradicts them. Weeks et al. (2016) conclude by saying that the idea that filter bubbles strongly reinforce confirmation bias is unlikely.

A second criticism of the filter bubble concept is that the algorithms learn users' preferences from their choices. Their choices can be expressed in different ways on social networks: through clicks, likes, shares and so on. On the basis of these different choices, the algorithms will suggest similar content, thereby reinforcing their interests and limiting the diversity of content (Pariser, 2011). The problem with this approach is that it is based on a form of technological determinism (Davies, 2018) and strong behaviourism (Graham, 2017). This assumption implies that people are active in their choices but passive in the face of the content they receive. Hansson and Grüne-Yanoff (2012) as well as Webster (2017) challenge this view by saying that

preferences are subjective mental states. They go on to say that choices are observable actions. Individuals may have multiple preferences influenced by social context, social desirability or heuristics. A “like” may be motivated by a desire to communicate a social image and not necessarily a genuine endorsement (Hart et al., 2020). Social networks are simply spaces where users put on a social show. They often click on popular articles regardless of the content (Messing & Westwood, 2014). Webster (2017) adds to all these assertions that the media also have their share of responsibility for users’ choices. In fact, not all choices come from users, given that the media push content via advertising. Dahlgren (2021) goes on to say that algorithms do not create a unique model for each individual. Just because we talk about personalisation does not mean that the result is personalised and that the algorithm really understands the person. The collaborative filtering we mentioned earlier is based on the behaviour of a large number of users, not just one individual. This filtering is therefore the opposite of the idea that each person is isolated in their own bubble. This aspect introduces a collective dimension into the recommendations. To conclude this criticism, we can state that an algorithm cannot deal with what is observable, given that it can only access an individual's internal preferences. Preferences are also multiple, contextual and often contradictory.

We can continue this analysis by indicating that the majority of users on Twitter/X share messages from people with similar opinions. But if we look at the interactions between users, we see that they are often exposed to divergent opinions. Sharing information through a retweet has more impact than commenting on it because it reaches more people (Dahlgren, 2021). However, exposure to information does not guarantee that it will be remembered. Vraga et al. (2016) explain that even after direct exposure, people quickly forget what they have seen. Sülflow et al. (2019) add that people read more when the content and responses present divergent opinions. According to Bakshy et al. (2015), users exposed to a personalised feed passively see more divergent opinions whereas users who actively choose, see fewer. Unlike them, Beam (2014) finds the opposite effect. This contradiction shows that personalisation can both reduce or increase exposure to diversity of opinion, but these changes are minimal. It is true that the changes are far from total exclusion as suggested by the filter bubble thesis. What we can learn from this is that audience fragmentation is rather an accidental side-effect of social platforms and not a direct consequence of algorithms. People interact with those who are different from them, sometimes out of preference and sometimes out of involuntary exposure.

Continuing this critical analysis of the filter bubble thesis, we see that it is based on the premise that Internet users now consume information solely via personalised services, often controlled by large, profit-driven American companies. This would make access to unfiltered information impossible. According to Pariser (2011), this personalisation creates an information environment unique to each individual, where certain important news items no longer appear because they do not receive enough interaction. The problem with this approach is that it focuses too much on a single medium. It is true that studying a single type of media (such as social networks, for example) without considering all media uses leads to an overestimation of the effects of filter bubbles. For example, a user might decide to find out about general news via traditional media and use social networks to follow specific interests (Dahlgren, 2021). This information has been proven by Fletcher and Nielsen (2017, 2018) who show that users consult various media and are thus exposed to a wide variety of (possibly contradictory) information. They add that even people who follow niche media consume mass media regardless of the country analysed. A. M. Guess et al. (2020) even contradict the conventional wisdom by stating that fake news does not spread massively on social networks. People hear about fake news first through the traditional media rather than on social networks (Tsfati et al., 2020). Most people do not consume information solely via social networks. They also visit news sites directly (Flaxman et al., 2016) or do not consult the news at all (Arceneaux & Johnson, 2013). So even if a user is locked into a bubble of filters on one platform, they can find divergent opinions elsewhere. Fragmentation does not just take place within a medium but also between media.

To conclude, the analysis of the algorithmic mechanisms of social platforms and the definition of the “filter bubble” concept in the previous sections have shown that the personalisation of the web creates unique and isolated information environments for each user. Consequently, this “bubble” phenomenon seems to exist within social platforms. However, this critical analysis reveals that certain nuances need to be added to Pariser’s theory.

Indeed, users look for information that confirms their beliefs, but they do not actively avoid contrary opinions. Algorithmic personalisation slightly reinforces this bias but does not create a total lock-in. What is more, social networks use algorithms that only interpret observable actions such as clicks or likes, but not real internal preferences. We also need to take into account the fact that information consumption is multi-channel (social networks, traditional media, etc.). Audience fragmentation is therefore more a side-effect than an intentional consequence of algorithms.

2.2.2. Fake news

2.2.2.1. Definition and classifications

One of the consequences of filter bubbles is to reinforce a user's existing beliefs and limit exposure to different points of view. These bubbles tend to create a favourable environment for the dissemination of misleading or false information (Flaxman et al., 2016; Pariser, 2011). In this context, one may question whether fake news could be a consequence of the personalisation of recommendations within social platforms. It is therefore essential to define what is meant by "fake news" in order to better understand how it spreads and how it differs from other types of misinformation.

The study of news veracity is not new, but it is only recently that the term "fake news" has been popularised and attracted widespread interest, following the US presidential election and Brexit referendum that took place in 2016 (Gelfert, 2018; Pennycook & Rand, 2021). So far, there is no universal definition for this term (Zhang & Ghorbani, 2020; Zhou & Zafarani, 2020). However, there are various characteristics of fake news that are generally accepted. In 2018, Gelfert defined the term "fake news" as "the deliberate presentation of (typically) false or misleading claims as news, where the claims are misleading by design" (p. 108), the expression "by design" referring to their systemic dimension. In their article, which has been widely cited in other papers, Allcott and Gentzkow (2017) explain that, for them, fake news corresponds to a news article that is intentionally and verifiably false and could mislead readers. Fake news is therefore deliberately erroneous or deceptive information presented in the form of a news. The author does this with the aim of acquiring a gain, which may be political or financial among other things (Meel & Vishwakarma, 2019; Zhang & Ghorbani, 2020). Meel and Vishwakarma (2019) add that these news pieces are usually published in news outlets (Zhou & Zafarani, 2020) or on the Internet (Zhang & Ghorbani, 2020).

Existing literature often refers to closely related terms when discussing fake news: misinformation, disinformation and mal-information. These concepts mainly differ in their intention and the veracity of the information presented. Misinformation corresponds to false information (Greifeneder et al., 2021; Meel & Vishwakarma, 2019; Pennycook & Rand, 2021). However, it is not circulated with the aim of deliberately causing harm (Greifeneder et al., 2021). According to several papers, the inaccuracy is generally due to an accidental mistake made by the author (Meel & Vishwakarma, 2019; Pennycook & Rand, 2021). On the contrary, disinformation is created with the deliberate intention of hurting someone (Greifeneder et al.,

2021; C.-C. Wang, 2020) or misleading the public (Meel & Vishwakarma, 2019; Pennycook & Rand, 2021; Zhou & Zafarani, 2020). Mal-information is also intended to cause harm (Greifeneder et al., 2021; C.-C. Wang, 2020). Nevertheless, what distinguishes the two practices is that mal-information is based on real facts, whereas disinformation spreads incorrect information (Greifeneder et al., 2021; C.-C. Wang, 2020).

While fake news is often deliberately created, it does not always come from a single type of source. They can be produced by individuals, organised groups or even governments, each pursuing different objectives: political influence, financial gain or ideological control (Meel & Vishwakarma, 2019).

With the rise of social media and digital platforms, it has become easier than ever for these actors to publish and disseminate misleading information on a large scale, and at lower cost (Zhang & Ghorbani, 2020; Zhou & Zafarani, 2020). Some creators launch into the production of sensationalist stories in order to attract attention, generate clicks and advertising revenue, without worrying about the accuracy of the content (Allcott & Gentzkow, 2017).

But creators are not the only ones contributing to the spread of fake news. Ordinary users also play a central role in their dissemination, often unintentionally. Stories are often shared on the basis of emotion, personal convictions or a desire to inform others, even if the content is incorrect (Greifeneder et al., 2021; Pennycook & Rand, 2021). According to Zhou & Zafarani (2020), the very structure of social media – where sharing is easy and responsibility limited – facilitates the virality of these misleading stories.

So, while fake news generally stems from a deliberate intention to deceive, its propagation is based on a complex combination of motivations, digital channels and actors, whether or not they are aware of the inaccuracy of the content.

In addition, it is important to note that the sources of information on the networks are extremely varied, and their reliability does not depend on their popularity. Influencers and youtubers are now competing with journalists and scientists, calling into question the role of traditional gatekeepers. This development is helping to blur the boundaries between producers and consumers of information. Finally, information overload increases exposure to rumours and reinforces their credibility through the familiarity effect (B. Kim et al., 2021).

2.2.2.2. Why do people believe fake news?

At a time when we are constantly bombarded with information online, it may seem surprising that so many people continue to believe fake news. Yet this phenomenon can be explained by

a combination of psychological, social and emotional factors. Between cognitive biases, the power of emotions, digital habits and trust in certain sources, the reasons are numerous and often unconscious. These reasons, together with brief explanations, are summarised in Table 2.

Researcher Kucharski (2016) offers an interesting comparison: he likens the spread of fake news to that of a virus. Social networks play the role of vectors, allowing rumours to spread, mutate and even compete with other news, just as different strains of virus do. This metaphor helps us to better understand why some false information goes viral.

Another paradox noted by Flanagin & Metzger (2000) is that even though many people say they distrust the information they find online, they tend to believe it all the same. In fact, certain characteristics of fake news (assertive tone, well-done visuals, closeness to our opinions) make it more convincing (Wogalter & Mayhorn, 2008). Even when a piece of information seems dubious, we can tell ourselves that it may contain some truth, and this is sometimes enough to give it weight (Kanoh, 2018).

Research into the sociology and psychology of the media has also shown that the way we consume news is strongly influenced by our personal preferences (Shu, Bernard, et al., 2019). For example, the concept of naive realism (Ross & Ward, 1996) refers to the tendency to believe that our view of the world is the correct one, the most “neutral”. And with confirmation bias, we naturally tend to go for information that confirms what we already think (Nickerson, 1998). This bias plays a central role in the way we receive fake news. If a piece of information is in line with our ideas, we are more likely to believe it (Nickerson, 1998), especially in politics (Michael & Breaux, 2021). That said, some researchers point out that it is not just a question of ideology: it is often our intuitive, quick reasoning that leads us to believe it (Pennycook & Rand, 2019). On the other hand, political opinions have a major influence on whether or not we share fake news (Pennycook et al., 2020).

We also need to talk about the illusory truth effect. It is simple: the more information we see, the more likely we are to believe it to be true (Pennycook et al., 2018). Even if we saw it the first time and knew it was false, repeating it creates a deceptive familiarity. This is even stronger if we are asked to judge whether a piece of information is ‘interesting’ rather than whether it is true (Calvillo & Smelter, 2020). And of course, if this information is relayed by sources we trust, the effect is amplified (Melki et al., 2021; Nadarevic et al., 2020).

Fortunately, certain factors can protect us: more analytical thinking, a good dose of critical thinking or sound general knowledge (Machete & Turpin, 2020; Pehlivanoglu et al., 2021; Pennycook & Rand, 2020; Sindermann et al., 2021). But analysing requires time and attention, two things we do not always have when we are scrolling through the news on our phone (Cordonier & Brest, 2021). What is more, some fake news is presented in a very professional way, with computer graphics, graphs, etc. which reinforces its credibility (Preston et al., 2021). Hence the importance of a good media education and a certain emotional intelligence (Salovey & Mayer, 1990).

Finally, the role of emotions should not be underestimated. Content that provokes anger, fear or anxiety attracts more attention and is shared more quickly (Berger & Milkman, 2012). The problem is that these emotions reduce our capacity for critical reflection. Emotions linked to analysis, on the other hand, do not have this effect, but if we encourage people to follow their feelings, this can increase their support for fake news (Martel et al., 2020).

If we believe fake news, it is not necessarily because of a lack of intelligence or education, but rather because of the way our brain, our emotions and our habits work. We are influenced by what we already believe, by what we feel, and by the way information is presented to us.

Reasons Why People Believe Fake News	Key Insights
Viral nature of content on social media	Fake news spreads like a virus, evolving and competing for attention just like real information.
Distrust yet belief paradox	Even when people claim to distrust online information, they may still believe it due to persuasive presentation.
Persuasive features of fake news	Assertive tone, polished visuals, and similarity to personal beliefs increase credibility.
Naive realism	People believe their perspective is objective, making contradictory info seem biased or false.
Confirmation bias	We naturally accept information that supports our existing beliefs, especially in politics.
Fast, intuitive thinking	Quick judgments often override critical thinking, leading to belief in false information.
Repetition effect (illusory truth)	Repeated exposure to the same false information makes it feel familiar and thus more believable.
Trust in the source	If misinformation comes from a trusted person or platform, we are more likely to accept it as true.
Emotional content	Content triggering fear, anger, or anxiety grabs attention and spreads quickly, while reducing reflection.

Table 2: Reasons why people believe fake news

2.3. Link between filter bubbles and fake news: existing studies and motivation for our research

Having explored the concepts of filter bubbles and fake news in greater depth, we can now review the papers that have examined these notions.

A number of recent studies in the scientific literature have investigated the circulation of fake news on social networks.

Pierri et al. (2020) tried to classify news articles on social media solely on the basis of the topological analysis of their diffusion networks. Their approach consists of an in-depth comparison of networks using non-alignment methods such as global network properties, distributions of centrality measures and distances between networks. The researchers found that the Twitter/X diffusion networks associated with reliable and misleading articles have different topological characteristics, which can be exploited to distinguish them automatically. In fact, they discovered that mainstream (reliable) and misleading (fake news, satire, rumours, clickbait) information generate diffusion networks with different structures. For example, these differences may relate to the size of the components, the depth of the cascades or the local density (clustering). Pierri et al. (2020) concluded that it is possible to automatically detect misleading articles on social networks simply by analysing the structure of their distribution on Twitter/X without analysing their content.

As for Stein et al. (2023), they sought to test the following hypothesis: “ideologically segregated networks do not simply favour the dissemination of partisan news but also disproportionately favour the dissemination of false information, particularly that which is too implausible to be propagated in integrated networks”. To do this, they set up controlled experiments with integrated or segregated artificial networks, the dissemination of true and false messages with an ideological bias and simulation models to generalise their results. The researchers observed that the participants in their study shared false partisan information at a similar rate in the integrated and segregated networks. There was no significant difference in individual willingness to share misinformation. However, in segregated networks, supporters are exposed to ideologically aligned false messages. This exposure is considerably reduced in integrated networks. The authors also noted that the proportion of false information circulating is higher in segregated networks, meaning that segregation structurally favours misinformation even if individual behaviour does not change. Messages with a low initial probability of being shared (low credibility) benefit most from segregation. For highly plausible messages, there is no difference between integrated and segregated networks. Authors’ general conclusion is that

ideological segregation in social networks greatly increases the exposure and dissemination of misinformation, especially when it comes to content with little credibility, even though individuals are no more inclined to share fake news themselves.

Finally, the study carried out in 2024 by Muñoz et al. aimed to gain a better understanding of the dynamics of disinformation and identify ways of fostering safer and more pluralist information. To do this, the authors carried out content and network analyses, creating network graphs to represent the interactions between users and thus the flow of information. These networks enable them to compare their evolution, identify the factors accelerating propagation, assess the robustness of the networks to propagation and study the correlation between the type or volume of content shared and the configuration of the network. The analyses performed revealed that the disinformation network is denser, more connected and more centralised than the legitimate journalists' network. These characteristics therefore enable better dissemination of information, suggesting the existence of coordinated structures aimed at influencing discourse. Muñoz et al. (2024) also noted a propensity to rely on content with a negative emotional charge to amplify their activity. Disinformation accounts will behave in a coordinated fashion. They are more effective at spreading information.

Several authors have also taken a more detailed look at the phenomena of filter bubbles and/or echo chambers, adopting various perspectives.

Related to what was discussed above regarding recommender systems, Nguyen et al. (2014) wanted to evaluate the long-term impact that a recommender system based on collaborative filtering can have on the diversity of content to which users are exposed, in order to measure the filter bubble effect at the individual level. To do this, the researchers employed a longitudinal approach to track changes in user experience over time. They also applied a measure of content diversity based on user-generated tags. Finally, they carried out a temporal analysis proposing new methods for observing the temporal effect of recommender systems on user experience. Through these different analyses, Nguyen et al. (2014) identified a reduction in diversity, with a moderate effect for engaged users. This is because recommender systems tend to expose users to a slightly smaller set of items over time. However, users who actually consume the recommended content are less affected by this reduction in diversity and rate items more positively.

Following a more algebraic approach, Lahoti et al. (2017) questioned the possibility of modelling and mitigating ideological filter bubbles on social networks (in particular, on Twitter/X) by automatically learning the ideological positions of users and information sources.

They modelled their research as a constrained non-negative matrix factorization, taking into account the structure of the social network (relationships between users) and content consumption data (user-content interactions).

Research has also been carried out into how different social media platforms influence the formation of echo chambers and the dissemination of information, particularly on controversial subjects. In their paper, Cinelli et al. (2021) seek to compare the social and informational dynamics between several platforms (Gab, Facebook, Reddit and Twitter/X). Their aim is to measure the impact of two factors: homophily in interaction networks and the bias towards disseminating information to users with similar opinions. Their study shows that Facebook and Twitter/X social platforms have a high degree of homophily in their interaction networks. This characteristic encourages polarisation and the formation of echo chambers where content circulates between individuals of the same ideological view. Cinelli et al (2021) also compare Reddit and Gab and find that users interact in more homogeneous communities with no clear separation between opposing groups. There is a degree of polarisation, but exchanges are not strongly compartmentalised. Echo chambers are less marked. This difference with Facebook and Twitter/X suggests a greater mix of audiences or a single dominant echo chamber. When the authors compare the Facebook and Reddit platforms on a same piece of content, Facebook shows strong segregation between users according to their ideological orientation. Reddit, on the other hand, shows more fluid interaction, even for users with extreme opinions, limiting echo chambers. In addition, Cinelli et al. (2021) note that recommendation algorithms play a central role. Facebook and Twitter/X offer little or no user control over their news feeds, increasing polarisation. Reddit allows more personalisation, limiting echo chamber effects. They conclude that the dynamics of polarisation and dissemination of information are highly dependent on the technical characteristics (algorithms, structure of interactions) specific to each social network.

More recently, Chueca Del Cerro (2024) tried to understand how social network configurations and the media environment influence the polarisation of national identity. To do this, she used an integrated model that corresponds to an agent-based simulation of national identity dynamics. It incorporates a multi-layer network of agents interacting in a hybrid environment combining traditional and social media. The author concluded that the initial structure of social networks (homophily based vs random) and the presence of filter bubbles play an important role in the dynamics and polarisation of ideas. Indeed, Chueca Del Cerro (2024) highlighted that the scenario combining a homophilic network and a filter bubble generates the strongest

polarisation, insisting on the fact that the presence of filtering algorithms on social networks reinforces polarisation.

However, few studies have examined the notions of filter bubble and fake news together.

A first study linking these two concepts is the research done by Zimmer et al. (2019). The authors aimed to understand the role of individuals and algorithms in the formation of filter bubbles and the support of echo chambers online, particularly in relation to the spread of fake news and deception on social media. To this end, studies of concrete information behaviour in relation to fake news and case studies with content analyses were carried out to discover interactions and links of cause and effect. As a result of these analyses, the authors discovered that social media algorithms do not directly create communities but amplify users' informational behaviour. They added that the key factor in the propagation of fake news and deception is the individual user, his cognitive patterns and his surrounding echo chamber.

Another study on the subject was done by Diaz Ruiz & Nilsson (2023). Although it does not look directly at filter bubbles, it studies the link between fake news and echo chambers. The authors investigated the way in which disinformation circulates on social networks through adversarial narratives integrated into identity-driven controversies. To do this, they set up an empirical analysis of the flat Earth echo chamber on YouTube. The main finding was that disinformation circulates via identity-based grievances and that the arguments put forward resist fact-checking because they are based on emotions and resentments, on a strong identification with the group and not just on objective facts. Diaz Ruiz & Nilsson (2023) also highlighted that disinformation circulates on social networks through a two-phase framework: first "seeding" and then "echoing". Disinformation then spreads within identity-based echo chambers fuelled by community involvement. Finally, the authors concluded that disinformation is not based solely on lies but also uses partial truths and emotional arguments. Social and cultural identities are at the heart of the spread of disinformation. Fact-checking alone is ineffective because it does not address the identity and community dynamics that anchor these beliefs.

Yet, the link between filter bubbles and the spread of fake news on social media deserves to be studied in greater depth. Several scientific studies suggest that a user enclosed in a bubble is more exposed to homogeneous content, and that this homogeneity may play a key role in the sharing of misleading content (Dahlgren, 2021; A. M. Guess et al., 2020). Del Vicario et al. (2017) have also shown that homogeneous communities on social networks are particularly

vulnerable to the circulation of false information, because the redundancy of narratives reinforces their apparent credibility. As a result, people may be led to share it, not because they believe it, but because they no longer have the necessary elements to exercise their critical faculties. It is also important to place user behaviour within the technical framework that makes it possible. Social platforms value rapid engagement: the more people react to a piece of content, the more it is promoted. Fake news, which is often emotional, polarising or sensationalist, fits this logic perfectly. But in a filter bubble, this mechanism becomes even more effective: content is calibrated to appeal to the target audience, which maximises its reach (Vosoughi et al., 2018). So, a user trapped in a bubble is not only more exposed to content that goes his way; he is also more encouraged to share it. Users do not necessarily see contradictory warnings, and they receive few dissonant signals (few critical comments, few fact-checks, little thematic diversity). This lack of friction facilitates impulsive sharing, often without verification (Pennycook & Rand, 2019).

However, it would be too simplistic to say that anyone trapped in a filter bubble automatically shares fake news. Recent studies qualify this view. Dahlgren (2021) points out in particular that exposure is not enough: some people exposed to fake news do not share it or recognise it as such. Similarly, A. M. Guess et al. (2020) show that individuals who are more analytical or more distrustful of networks are less likely to relay fake news, even if they are exposed to it. This is why it is relevant to look in more detail at the relationship between the sharing of fake news and enclosure in a filter bubble on social networks. To the best of our knowledge, this link has only been studied artificially in the literature so far. Through the study of empirical data we collected, we will seek to determine whether in practice a link can be established between these two phenomena. Furthermore, still based on this data, we will also try to learn more about the behaviours that characterise users who spread fake news online.

3. Link between filter bubbles and the sharing of fake news: a practical approach

3.1. Methodology

3.1.1. Research questions to be explored and approach used

As highlighted in the literature review, we are seeking to understand whether there is a link between filter bubbles and the spread of fake news on social networks. We also intend to learn more about the behaviour of users who share fake news online. More specifically, we aim to answer the following research questions:

- **RQ1:** Are users who share fake news on social media more likely to be trapped in a filter bubble?
- **RQ2:** Do users who share fake news on social media tend to have stronger opinions?
- **RQ3:** Do users who share fake news on social media tend to share more recommended content, such as retweets or quoted tweets, rather than original tweets?
- **RQ4:** Are there any explicit features that distinguish users who share fake news on social media?

For this study, we will adopt a quantitative approach. Quantitative analyses are well suited to testing theoretical predictions using measures (Tharenou, 2007) such as mathematical models and statistical tests. This type of approach will enable us to concretely measure conceptual phenomena such as diversity of content, emotional intensity of opinions or publication behaviour on empirical data, and subsequently to compare two groups of users.

For our analysis, we will base ourselves on the social network Twitter (newly named X). Twitter/X is widely used to share information in real time on a wide range of topics, including news, politics and entertainment, which are often subject to misinformation. Moreover, due to its operating mechanisms and the behaviour patterns of its users, Twitter/X is a platform that is propitious to the development and propagation of misinformation and disinformation (Muñoz et al., 2024; Starbird et al., 2019). With around 375 million tweets posted every day (Pfeffer et al., 2023), the platform is proving relevant for observing the circulation of fake news.

In addition, Twitter/X has become a popular research medium in recent years, and various problems have been analysed using data from the platform (Chen et al., 2021). Although the rules governing the use of Twitter/X's API have recently been tightened, making data collection

more laborious, the broad range of targeted data that can be extracted is valuable for studying the phenomena we are interested in.

3.1.2. Data collection

In order to study our different research question, we based ourselves on articles analysed by the website Gossip Cop. We do this because our different research questions require us to consider the veracity of some news items circulating on Twitter/X. Gossip Cop was a New York-based website that checked the veracity of entertainment news, mainly celebrity-focused, published in various media such as magazines, newspapers and on the internet. Before providing detailed explanation, Figure 3 summarises all the steps performed to collect the data needed for our analyses. The various codes used to reproduce our work can be found on Github¹.

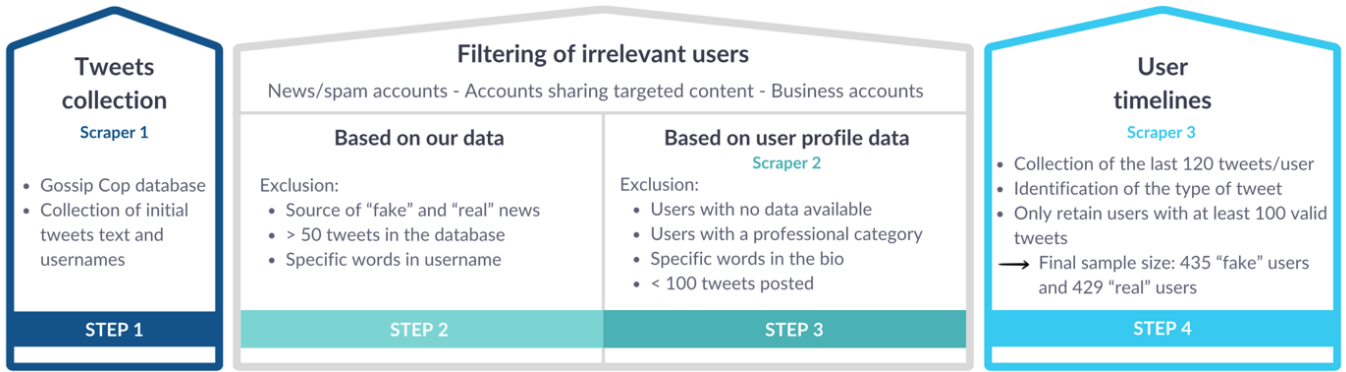


Figure 3: Summary of the data collection phase

The starting point for our data collection is a database containing a wide range of Gossip Cop news items (Shu, Mahudeswaran, et al., 2019; Shu, Sliva, et al., 2017; Shu, Wang, et al., 2017). As the examples in the table below show (Table 3), for each article we have information on its origin, its title, a label that characterises the veracity of the story (either "fake" or "real"), and a list of tweet ids that refer to this news item on Twitter/X.

id	news_url	title	tweet_ids	label
gossipcop-2493749932	www.dailymail.co.uk/tvshowbiz/article-5874213/Did-Mil...	Did Miley Cyrus and Liam Hemsworth secretly get married?	284329075902926848 2843327445599...	fake
gossipcop-941805037	variety.com/2017/biz/news/tax-march-donald-trump-...	Celebrities Join Tax March in Protest of Donald Trump	853359353532829696 8533595765439...	fake
gossipcop-882573	https://www.brides.com/story/teen-mom-jenelle-evans-...	Teen Mom Star Jenelle Evans' Wedding Dress Is Available ...	912371411146149888 912371528343408...	real

Table 3: Extract from the "news" table in the database

¹ [GitHub - emmagarcia08/fake-news-and-filter-bubble-analysis](https://github.com/emmagarcia08/fake-news-and-filter-bubble-analysis): Python scripts used in our master's thesis, which examines the link between the sharing of fake news and the filter bubble phenomenon on Twitter/X.

We conducted an initial collection phase. For articles that were still available, i.e. those that had not been deleted and for which we were able to retrieve the original text, we collected the tweets that referred to them using their tweet id. To do this, we used a scraper² from the Apify website. Apify is a third-party platform that offers scraping tools and other solutions designed by developers. For each tweet still available on Twitter/X, we collected its text and the username of the user who posted it. We also kept track of all the other information related to the tweet returned by the scraper. This information is stored in tables named “collected_tweets_fake” and “collected_tweets_real” in the database. The Table 4 is an example of what these tables contain.

tweet_id	username	text	full_data
1033430094533066753	DailyMailCeleb	Cheryl 'split from ex Liam Payne as she felt like she was ...	{ "tweet_id": "1033430094533066753", "lang": "en", ...
835210670962577408	kasc4u	Khloe Kardashian NOT Begging Tristan Thompson To ...	{ "tweet_id": "835210670962577408", "lang": "en", ...
844534701486354434	_coinmerch_	President Donald Trump and First Lady Melania Keep ...	{ "tweet_id": "844534701486354434", "lang": "en", ...

Table 4: Extract from the “collected_tweets_fake” table in the database

Although our resources and access to the scraper were limited, we managed to collect a total of 353,654 tweets. Specifically, 167,654 tweets refer to 3,359 articles identified as “fake”, and 186,000 tweets refer to 9,920 articles identified as “real”.

Now that we have identified the users behind the tweets referring to articles labelled as real or fake, the next step is to analyse more precisely the content to which these users are exposed. However, not all accounts listed in the “collected_tweets_fake” and “collected_tweets_real” tables are suitable for our analysis. Our goal is to study the filter bubble phenomenon among ordinary users, without any particular distinction. However, some Twitter/X accounts in our database are by nature intended to relay specific content. These accounts may have low content diversity, leading us to believe that they are trapped in a filter bubble when in fact they are deliberately aiming to share targeted content with their community. Including them in our study would therefore risk introducing bias, artificially lowering diversity scores. We must therefore exclude these accounts as much as possible in order to avoid drawing misleading conclusions. The users to be excluded can be grouped into three main categories. The first group concerns news accounts. These are accounts that aim to inform their community about current events,

² https://apify.com/coder_luffy/free-tweet-scraper

relay star gossips or share entertainment. Also in this category are ‘spam’ users, who share continuously catchy tweets in an identical format (see Appendix 1). Given that our study is based on tweets about a diversity of articles, these news accounts are well represented among the users we have collected. The second category of accounts concerns users who share targeted content. This includes people who publish tweets on a specific topic, or fan accounts that provide information about a particular public figure. Finally, the last type of user to be discarded are business accounts. These are professional accounts through which people promote their business.

In order to exclude these users, we carried out an initial filtering based on their occurrence in our data and the analysis of usernames.

First of all, we eliminated users who were the source of both “fake” and “real” tweets, leaving only people associated with just one of the two labels. We also ignored people who had published more than fifty tweets among those we collected. This is because, when analysing various users, we noticed that those who appear more than fifty times in the data are generally among the users who are of no interest to our study. For example, this operation allows us to eliminate some of the news accounts, which by their very nature are the authors of a large number of these tweets.

We then filtered the remaining people on the basis of their username and the name displayed on their profile. We removed users whose name contained one of the following words: news, media, info, radio, celebrity, fan.

This initial filtering allowed us to retain a total of 93,695 users, i.e. 77,956 users linked to fake article(s) and 15,739 users linked to real article(s). These users were stored in a table called “users”.

Among the remaining people, we noticed that irrelevant users could still be filtered on the basis of some characteristics in their profile. We therefore collected user data from all the remaining accounts using a second scraper³ found on Apify. As before, we saved all the information about the users returned by the scraper, but we were mainly interested in the text of their bio and whether or not they had a professional category in their profile. We stored these elements directly in the “users” table, as shown in Table 5. As the bios are not all in English, we have also translated them using the deep_translator package⁴ via a python script, so that all users can be filtered on the same basis.

³ <https://apify.com/kaitoeasyapi/premium-twitter-user-scraper-pay-per-result>

⁴ <https://deep-translator.readthedocs.io/en/latest/index.html>

username	label	bio	professional_category	full_data	bio_translation
00Zixx	fake	Simply Ill Clothing link in bio	NULL	{"action_counts": {"favorites_count": 7397}, "avatar": ...}	Simply Ill Clothing link in bio
0131chaez	real	슬픈 거북이	NULL	{"action_counts": {"favorites_count": 2037}, "avatar": ...}	Sad turtle
OnePuff	fake	Mountain of Peace Shrine Founder, Catholic, Former Political Prisoner, Writer and Smoking Cessation	Social Media Influencer	"action_counts": {"favorites_count": 1304}, "avatar": ...}	Mountain of Peace Shrine Founder, Catholic, Former Political Prisoner, Writer and Smoking Cessation Therapist in One Virtual Puff.

Table 5: Extract from the “users” table in the database

We then carried out a second filtering based on these new data. We eliminated the people for whom the scraper was unable to recover anything. This means that their Twitter/X account is no longer available. We also removed all users who indicated a professional category in their profile. This makes it possible to exclude some of the corporate accounts and accounts that aim to promote a business, which are more complicated to filter on the basis of other criteria.

Of the remaining people, we excluded those with one of these words in their bio: news, media, info, radio, station, entertain, celebrity, celebrities, fan. Finally, we also ignored accounts that had shared less than 100 posts in total. This is to ensure that we only retain users with a minimum amount of content to analyse later.

This second filtering phase gave us a final total of 72,989 users, with 61,473 people labelled “fake” and 11,516 labelled “real”. Obviously, some users who were not relevant to our study could not be filtered out and were still included in the latter list, but the two filtering phases nevertheless enabled us to exclude a total of 33,409 Twitter/X accounts.

Finally, we collected the latest tweets from users identified as “relevant”, in order to analyse their content later on. To do this, we used a third scraper⁵ from the Apify website, which enables us to collect a user's timeline. As our resources were limited, we decided to restrict our collection to a set of 1,000 users from among those remaining after filtering, with 500 users having shared a fake article and 500 users having shared a real article. For each of these users, we asked the scraper to collect their last 120 tweets. Although users have generally shared a large number of posts, some tweets are not always accessible. This ensures that sufficient content is retrieved for most users, so that a fairly representative analysis of their diversity can subsequently be conducted.

We have stored the information gathered by the scraper in a new table (see example in Table 6). For each tweet collected, we retrieve the username it comes from, the label associated with

⁵ <https://apify.com/danek/twitter-timeline-ppr>

that user (“fake” or “real”), the tweet id and its type (“tweet”, “retweet” or “quoted_tweet”). A “tweet” is a post written directly by the user of the account being analysed. If it is a tweet posted by another user that the analysed user has shared with his or her followers, we call it a “retweet”. Finally, a “quoted tweet” is a retweet for which the analysed user has added a comment to add his or her opinion to the original tweet that was written by the other user. The “text” column of the table shows the text of the tweet in the case of a tweet or retweet. In the case of a quoted tweet, the “text” column contains the comment posted by the analysed user and the text of the original tweet that has been reposted can be found in another column called “quoted_text”. This enables us to retain the context of the user's comment, as the comment alone is generally difficult to interpret. We then translated the texts in these two columns into English as before, so that all users could be analysed in the same way.

username	label	tweet_id	text	quoted_text	tweet_type	full_data
09TaLEnTd18	real	681655444604108800	Waiting on a room	NULL	tweet	{"tweet_id": "681655444604108800", ...
1BigFlizzy	fake	1401732306017132545	Paul Archibald	Yo who is this announcer lol	quoted_tweet	{"tweet_id": "1401732306017132545", ..
Anajonas94	real	1687544435007954944	Achievement breeds opportunity.	NULL	retweet	{"tweet_id": "1687544435007954944", ..

Table 6: Extract from the “user_timelines” table in the database

With regard to collecting the content of tweets, we encountered two limitations. For some tweets, the scraper was unable to retrieve the text because it was not accessible to us for legal reasons. Also, some tweets were recovered by the scraper but it was unable to retain any information about them, not even their id. This is because these tweets have a very specific format in the user's timeline. As a result, we have not taken these two cases into account in our data, as the content of the concerned tweets is missing. We also excluded tweets consisting solely of one or more URLs. This is because these web addresses alone do not allow us to learn more about the content consumed by the user.

Finally, of all the users for whom we were able to collect valid tweets, we chose to keep only those for whom we had at least 100 tweets, in order to have enough content to analyse per person. Although we have only considered users who have shared at least a hundred posts, some of them have a timeline available on their profile that is very small, or even inaccessible in some cases. For example, one of the users only has seven posts in his/her timeline, even though it says at the top of his/her profile that he/she has shared 7,105 posts (see Appendix 2). This is

why the scraper recovered a fairly low number of tweets for a number of accounts. Out of 1,000 users processed by the scraper, we are left with a total of 864 users who satisfy the condition of adequacy, subdivided into 435 users linked to the “fake” label and 429 users linked to the “real” label. We stored these users in a final table called “filtered_user_timelines” in the database, which we will use directly for our analyses.

Now that we have collected the data needed for our analyses, we will examine in detail each of the research questions presented above. We will develop why it is relevant to study them in the context of our thesis, and explain in depth the various processing steps we have applied to our data in order to answer our hypotheses.

3.1.3. Calculation of thematic diversity scores

Our main research question addresses the link between filter bubbles and the sharing of fake news (**RQ1**). As highlighted in the literature review, several authors suggest that exposure to homogeneous content – which results from being enclosed in a filter bubble – could play a role in the sharing of false information. Exploring this first question will therefore enable us to evaluate whether being enclosed in a bubble can lead to the propagation of misinformation.

The research question involves determining whether or not the users are trapped in a filter bubble. The first part of this thesis highlighted the fact that users trapped in a filter bubble are faced with more restricted, less diverse content. Consequently, we assume that a user is trapped in a filter bubble if this person is confronted with content that is low in diversity. We will therefore analyse the content faced by each of the users we have collected on Twitter/X and assess their diversity. When it comes to the content a user is confronted with, it is obviously not possible to track every tweet the user has ever encountered. However, the user’s Twitter/X timeline (i.e., the history of all the tweets shared by the user, presented in reverse chronological order, located under the “Posts” tab on the user’s profile page (see Appendix 3)) already gives a good idea of the content the user has been exposed to. We will consequently evaluate the diversity of each user’s timeline we collected following the data collection phase. For each user, the analysis of his/her diversity will be based on around a hundred tweets that he/she has tweeted, retweeted or quoted. It is worth noting, however, that a timeline with low diversity does not necessarily imply a lack of exposure to diverse content overall, as users may still actively seek out diversity on their own, without necessarily sharing this other content. Our analysis therefore captures the diversity of content users are confronted with that is actually

observable, which is an indicator of exposure to information, but does not take into account exploratory behaviour beyond the public timeline.

Given that we want to compare the filter bubble phenomenon between users who have shared real news and those who have shared fake news, we need to assign a diversity score to each of these users. To this end, we based ourselves on the method developed by Bache et al. (2013). Their approach is designed to assign a thematic diversity score to a document – in our research, the thematic diversity of a user represented by a group of tweets – on the basis of its textual content. As highlighted by these authors, the question of quantifying diversity has been studied in several fields, but few studies have focused on measuring diversity related to textual data. This is why the approach of Bache et al. (2013) proved to be very relevant to our study and easily transposable to our data. In addition, the authors use the measure developed by Rao (Rao, 1982), and later reused by Stirling (Stirling, 2007), to obtain diversity scores. This measure is generally considered to be comprehensive, since it takes into account the three components that characterise diversity: variety, balance and disparity (Leydesdorff, 2018; Mutz, 2022; van Dam, 2019).

The method of Bache et al. (2013) can be divided into three steps. Firstly, they train a Latent Dirichlet Allocation (LDA) topic model with collapsed Gibbs sampling on a corpus of documents – in our case, all tweets collected from all users – to form a document-topic count matrix. Secondly, on the basis of this matrix, they calculate the similarity between topics based on their co-occurrence in the documents, using cosine distance. These similarities are then transformed into distances between pairs of topics. Finally, they calculate a diversity score per document – in our analysis, per user – by applying Rao's diversity measure. In this section, we will develop in detail the method applied to our data. Firstly, a pre-processing step was performed on the content of the tweets and, based on the tokens obtained, a document-topic count matrix was constructed. Secondly, co-occurrence similarity calculations between topics were performed. Finally, based on these results, a diversity score for each user was found. The figure below (Figure 4) visualizes and summarizes the method applied to our data.

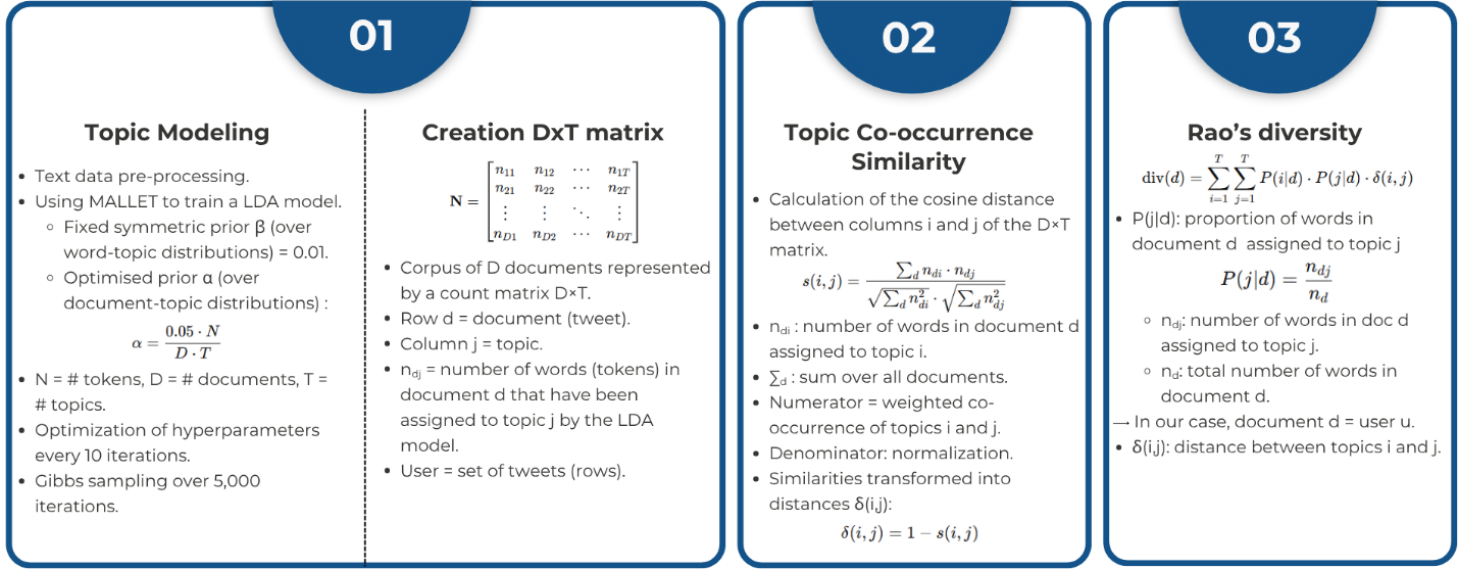


Figure 4: Summary of Bache et al.'s method applied to our study

3.1.3.1.Pre-processing stage

The first step to be implemented is a pre-processing stage to obtain tokens that can then be analysed.

When linguistically pre-processing tweets, we rely on NLP libraries such as Stanza and NLTK to clean, lemmatize and filter the texts.

First, a cleaning step is necessary within the collected tweets. We first put all the words in lower case. Next, we remove all URLs from tweet texts. This step makes sense for our analysis, given that the scraper collected URLs in their simplified form. No words can be used for our subsequent analysis, as the addresses contain only numbers and letters with no real meaning if we keep them. Then, we also remove non-alphanumeric characters.

Secondly, we set up the processing step to apply tokenization and lemmatization with stanza. Stop words, numbers and very short tokens are also filtered out. The result is a string of lemmatized tokens separated by spaces.

3.1.3.2.Creation of the document-topic count matrix

Once the content of the tweets is ready for use, we create the document-topic count matrix as done in Bache et al. (2013). To obtain it, a corpus of D documents is represented as a count matrix such as $D \times T$. In this matrix, each row d corresponds to a document – a tweet in the context of our analysis – and each column j represents a topic. A user is thus represented in this matrix by a set of rows containing the tweets he/she has written or shared. The entry (d, j) indicates the number of words (tokens) within a tweet assigned to a given topic.

To associate words with topics, a latent Dirichlet allocation (LDA) model with collapsed Gibbs sampling is used to learn the topics. We decided to use the LDA model rather than more recent approaches based on large-scale language models (LLMs) – which consist of a foundation model trained on huge amounts of self-supervised textual data, designed to be adapted to a wide variety of downstream tasks (Bommasani et al., 2022) – since these models present several limitations and risks in the context of our research.

Firstly, LLMs are opaque and difficult to interpret, making it challenging to reliably and explainably extract the thematic distributions (Bommasani et al., 2022) required to measure informational diversity. Conversely, LDA is a well-founded probabilistic model that represents each document as an explicit distribution on latent topics (Blei et al., 2003), which makes it more suited to our objective.

In addition, LLMs can reproduce or amplify the biases present in their training data while posing risks of misuse (e.g. disinformation campaigns) (Bommasani et al., 2022). These issues are particularly sensitive in a study dealing specifically with the propagation of fake news. In addition, LLMs require considerable computational resources and are often only accessible via proprietary platforms, which limits their transparency and reproducibility (Bommasani et al., 2022). In contrast, LDA is lightweight, open source and fully controllable, guaranteeing the independence of the research (Blei et al., 2003).

Finally, LDA can be used to model multi-theme documents (Blei et al., 2003), which is essential for analysing the thematic diversity of tweets from the same user. LDA is therefore a reproducible and well-established method for analysing large social network databases. For all these reasons, we consider that LDA is a more suitable solution for our study.

The LDA model learns on the basis of different hyperparameters. The first is β , which is the prior over the topic-word distributions (George, 2015). It is symmetric, and set to 0.01 as done in the method we draw on Bache et al. (2013). The second hyperparameter is the prior α , which corresponds to the prior probability distribution of the topic-document distribution (Du, 2022; George, 2015). It is initialized based on the formula $0.05 \frac{N}{D \cdot T}$, where N corresponds to the number of tokens in the corpus (= 1,214,706 tokens in our corpus), D is the number of documents (= 99,422 tweets in our corpus) and T is the number of topics. In order to determine the most appropriate number of topics based on our data, a topic coherence measure is applied for different topic values. We use the C_V coherence measure as it has been empirically validated in the literature as being correlated with human assessments of topic quality. It enables

automatic, robust and interpretable evaluation of topic models. This measure is a combination of the Boolean sliding window, normalized pointwise mutual information (NPMI) and indirect cosine measure (Röder et al., 2015). As we can see from the accompanying graph (Figure 5), the number T of topics should be equivalent to 30, since we obtain a better coherence score there. The value of parameter α is therefore equivalent to $0.05 \frac{1,214,706}{99,422 \times 30} = 0.02$. Optimization of these hyperparameters is activated every 10 iterations.

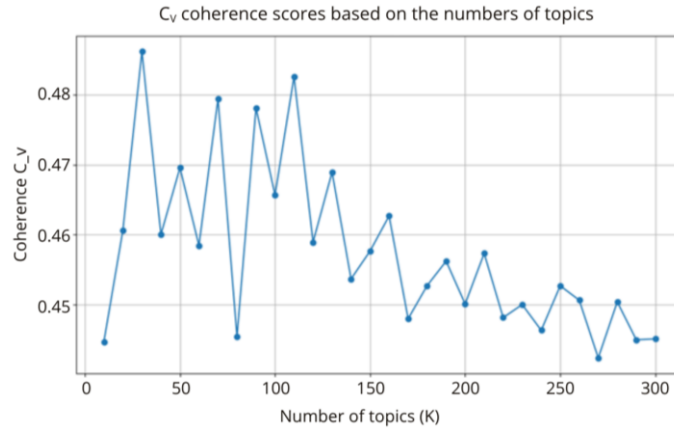


Figure 5: C_v coherence scores based on the numbers of topics

An iteration of the collapsed Gibbs sampler traverses all the tokens in the corpus and assigns a topic to each of them. Each sampler is run for 5,000 iterations, retaining only the final sample. At the end of the sampling, a $D \times T$ matrix is constructed, with n_{dj} corresponding to the number of tokens in document d assigned to topic j .

If we look at the quality of the tweet-topics assignments, we can note that it is good overall. The table below (Table 7) shows some examples of tweets from our database and the topics assigned to them by the model. For example, the first tweet in the table has been assigned to topics 5 and 11. Looking at the most representative words for each topic (see Appendix 4), we can see that topic 5 is about contests, prizes, etc. and topic 11 is about sport. Indeed, the author of the tweet suggests offering beers depending on the score of a football match.

Tweet text	Assigned topic(s) and their meaning
For every goal scored in tonight's Argentina v Nigeria game we will giveaway a World Cup beer glass. Retweet & follow to enter. #WorldCup #ARG #NGA	Topic 5: contests Topic 11: sport
Mike Johnson signals he will enact Trump's Project 2025 if they win in November: We're gonna make big, radical changes. Those first 100 days, aggressive changes to the country	Topic 18: US politics
I have crackers, but I'm out of peanut butter, the crackers have a great crunch, and they go well with peanut butter, so I guess I have to shopping for peanut butter so I can't eat my crackers with the great crunch!!!	Topic 4: food
Biometric concentration camps marketed as aid hubs, all run by US mercenaries and paid by Israeli intelligence via a Swiss shell company founded by the biggest names in humanitarian aid – including the former CEO of World Central Kitchen	Topic 0: business/money Topic 3: Israeli-Palestinian conflict
Welcome to the privatized Gaza genocide	Topic 18: US politics

Table 7: Examples of tweet-topics assignments

3.1.3.3. Topic co-occurrence similarity

To obtain diversity scores based on the formula presented by Rao, we need an inter-subject similarity measure to find the distance values between each pair of topics on which the diversity calculation depends.

In their research, Bache et al. (2013) identified that the similarity measure between two topics that provides the best results is the cosine similarity, based on the co-occurrence of subjects in documents (D x T matrix). We therefore decided to apply this similarity to our D x T matrix with columns i and j. The similarity cosine calculation is:

$$s(i, j) = \frac{\sum_d n_{di} n_{dj}}{\sqrt{\sum_d n_{di}^2} \sqrt{\sum_d n_{dj}^2}}$$

where n_{di} is the number of words in tweet d assigned to subject i.

3.1.3.4. Rao's diversity measure

The measure of thematic diversity of a document d is defined by Rao's formula such as:

$$div(d) = \sum_{i=1}^T \sum_{j=1}^T P(i|d) P(j|d) \delta(i, j)$$

In our study, we use this measure to calculate a diversity score for each user. Therefore, in this formula d does not correspond to a single tweet (seen as a document in the matrix of document-topic counts) but to all the tweets assigned to the user. From this perspective, $P(j|d)$ corresponds to the proportion of tokens in the user's tweets assigned to topic j, estimated by $\frac{n_{dj}}{n_d}$ and $\delta(i, j)$ represents the distance between topics i and j. To convert similarity $s(i, j)$ into distance $\delta(i, j)$, we use the transformation $\delta(i, j) = 1 - s(i, j)$. This transformation makes sense with our data and allows us to keep a final diversity score between 0 and 1, which is easy to interpret.

3.1.3.5. Method used to answer the research question

Once the diversity score has been found for each user, it is interesting to interpret these results to answer **RQ1**. We aim to test the following hypothesis: “users who share fake news on social media are more likely to be trapped in a filter bubble”. Firstly, we are going to perform a descriptive analysis of the sample. We will look at the distribution of diversity scores for the two groups of users (“real”/“fake” users). Statistical calculations will be carried out to see if there is a central tendency and dispersion within the two groups. A statistical analysis will also

be applied to test our hypothesis. After showing that the data are not normally distributed, the Mann-Whitney test will be used to compare diversity scores between users associated to real or fake news. The descriptive analysis of the thematic diversity scores can be found in the “Experimental results” section.

3.1.4. Calculation of sentiment scores

While the first part of our analysis focuses on content diversity as a structural indicator of the filter bubble effect, we now turn our attention to the possible consequences of this information lock-in on user attitudes. Indeed, the literature review highlighted that repeated exposure to homogeneous content can not only reduce information diversity, but also reinforce user’s opinions. That is why in a second phase, we will analyse whether users who share fake news tend to have stronger opinions (**RQ2**). By “stronger opinions”, we mean views that are more marked, positively or negatively, and more distant from neutrality. We are then referring to emotional intensity.

To investigate this, we carried out a sentiment analysis of tweets from all users. We followed the method described by Pandarachalil et al. (2015), which is well suited to our study. It consists of a lexicon-based approach, adapted to the processing of tweets, which is unsupervised and domain-independent. The authors try to assign a polarity score to the terms (four-grams, tri-grams, bi-grams and uni-grams) present in the tweets, based on three lexical resources. Unlike most machine learning-based approaches, it does not require tweets to be annotated, and its accuracy does not depend on a target domain (Y. Wang et al., 2022). However, the drawbacks of lexicon-based methods are that they do not take context into account and some words may not be found in the lexicon used (Y. Wang et al., 2022). Nevertheless, the work of Pandarachalil et al. tends to reduce the impact of these limitations by considering n-grams in the evaluation of the polarity of tweets, rather than just isolated terms. In addition, it broadens the scope of the terms being considered by using several lexical resources with varied content.

3.1.4.1. Pre-processing stage

Firstly, we pre-processed the tweets in a similar way to what is done in Pandarachalil et al. (2015). This step is essential before proceeding with the sentiment analysis, given that the tweets are often noisy and presented in an informal manner. We removed URLs, hashtags and mentions starting with the “@” character and replaced word contractions with their extended form. We then tokenised the tweets. To do this, we used the TweetTokenizer included in the

nlTK package⁶, which is specifically adapted to the processing of tweets so as not to eliminate elements that are crucial to our analysis, such as emojis and abbreviations. We also processed elongations by reducing repeated characters until we had a meaningful word. We then proceeded to the lemmatisation with WordNet using the WordNetLemmatizer included in the nlTK package⁷, except for the slangs, which we kept intact. One difference from the work of Pandarachalil et al. is that we also took care to keep the emojis and emoticons present in the tweets, since we believe they are strong indicators of emotion that should not be ignored. Finally, we also replaced words preceded by the negations “not” or “never” with an antonym where possible, using WordNet. We applied this pre-processing phase to the English translation of each tweet (text_translation column in the filtered_user_timelines table of our database), to ensure that we were always starting from a usable input. At the end of these various operations, the tweets were converted into a series of tokens, which we then transformed into four-grams, tri-grams, bi-grams and uni-grams in order to perform the sentiment analysis.

3.1.4.2. Calculation of a sentiment score per tweet

To assign polarity scores to the terms contained in the tweets, Pandarachalil et al. (2015) use three lexical resources: SenticNet for analysing n-grams (four-grams, tri-grams, bi-grams, uni-grams), SentiWordNet as a complement for analysing uni-grams, and SentislangNet for analysing slangs and acronyms. We followed a similar logic, but with a few differences. As SentislangNet was not made publicly available by the authors, we used the VADER lexicon from the vaderSentiment package (Hutto & Gilbert, 2014), which contains the polarity score of slangs as well as emoticons and uni-grams, thus enriching our analysis. To process n-grams, we used the latest version of SenticNet, i.e. SenticNet 8 (Cambria et al., 2024). We downloaded the lexicon⁸ and scraped its python script to create a SenticNet dictionary in JSON format that is easy to use. The latest version of SenticNet contains over 300,000 concepts, including emoticons and a wide range of emojis, allowing us to extend our analysis compared with the study by Pandarachalil et al. (2015). Finally, we also used SentiWordNet (Baccianella et al., 2010) to analyse uni-grams.

For each tweet, we assigned a sentiment score from its list of n-grams. We began by searching for a polarity score for four-grams, tri-grams and bi-grams via SenticNet. Then we tried to assign polarity scores to uni-grams, searching successively in SentiWordNet, then SenticNet

⁶ https://www.nltk.org/_modules/nltk/tokenize/casual.html

⁷ https://www.nltk.org/_modules/nltk/stem/wordnet.html

⁸ <https://sentic.net/downloads/>

and finally VADER. To avoid multiple counts of identical terms, polarity scores are not taken into account if the term (uni-gram, bi-gram or tri-gram) is contained in a n-gram already taken into account previously. Finally, an overall sentiment score is assigned to the tweet using this formula:

$$sent_score = \frac{\sum_{t \in term} pos_t - neg_t}{|term|}$$

where pos and neg represent respectively the positive and negative polarity score assigned to each term based on the three lexicons and term corresponds to the total number of terms taken into account in calculating the sentiment score. Regarding polarity scores (see Table 8), SentiWordNet returns two scores per uni-gram: a positive score (pos) and a negative score (neg), both within the interval [0,1]. SenticNet and VADER, on the other hand, associate a single score with each term, representing positive or negative polarity. Consequently, the second score is assumed to be 0 in the global sentiment calculation. For example, if SenticNet returns a positive score for the term, the negative score associated with the term will be defined as zero, and vice versa. In terms of value ranges, the score returned by SenticNet is within [-1,1], while the score assigned by VADER is within [-4,4]. Consequently, we applied a normalisation to the scores returned by VADER, so that these scores are contained within [-1,1]. This ensures that there are no differences between scores from different lexicons. The overall sentiment score of the tweet is a value between -1 and 1. A negative score means that the tweet is globally associated with a negative sentiment, whereas a positive score reflects a positive sentiment overall.

LEXICON	RAW LEXICON SCORE	REVISED SCORE
 SentiWordNet	pos score $\in [0,1]$ AND neg score $\in [0,1]$	pos $\in [0,1]$ neg $\in [0,1]$
 SenticNet	score $\in [-1,1]$	if score $\in [-1,0]$ pos = 0 neg $\in [0,1]$ pos $\in [0,1]$ neg = 0
 VADER	score $\in [-4,4]$	if score $\in [-4,0]$ pos = 0 neg $\in [0,1]$ pos $\in [0,1]$ neg = 0

Table 8: Comparison between the raw score returned by each lexicon for a term and the associated revised score

3.1.4.3. Calculation of a sentiment score per user

In order to analyse our second hypothesis, i.e. “users who share fake news tend to have stronger opinions”, we need a score that is representative of each user’s opinions. Since we are interested

here in the intensity of the user's sentiments and, more specifically, his deviation from neutrality, we have chosen to take the average of the absolute value of the sentiment scores associated with each user's tweets:

$$user's\ score = \frac{1}{N} \sum_{i=1}^N |score_i|$$

where N is the total number of tweets analysed for the user and $score_i$ is the sentiment score associated with tweet i. Each user will therefore have a score between 0 and 1. The closer the score is to 1, the more emotionally polarised the user is (very negative and/or very positive opinions). On the other hand, if the score is close to 0, this means that the user shares tweets that are generally more moderate (rather neutral opinions).

3.1.4.4. Method used to answer the research question

Once a sentiment score has been found for each user, we seek to confirm or refute the existence of a more pronounced emotional intensity among users who have shared fake news. In a similar way to the analysis of diversity scores, a descriptive and statistical analysis will be carried out on the sentiment scores obtained in order to answer our research question. This analysis can be found in the "Experimental results" section.

3.1.5. Method used to analyse user publication behaviour

Following these first two research questions, we are also interested in the influence that the platform's algorithm can have on user behaviour. We can ask ourselves whether people who share fake news are more sensitive to the content that is recommended to them, which could potentially explain why they end up sharing false information. To explore this question, it is relevant to look at the types of tweets users are sharing. On the one hand, original tweets published directly by the user can be seen as self-initiated content, resulting from the user's autonomous thinking. On the other hand, non-original tweets reposted by the user, such as retweets and quoted tweets, can be seen as received content, potentially influenced by Twitter/X's algorithm. Consequently, the proportion of retweets and quoted tweets shared by the user can be considered as an indirect indicator of algorithmic influence. This leads us to the third research question (**RQ3**). We will test the following hypothesis: "users who share fake news are more sensitive to content that is recommended to them, and therefore tend to share more relayed content (retweets/quoted tweets)". To do this, we will calculate the proportion of original and non-original tweets among the tweets collected for each user, and analyse the

distribution of these proportions for the two categories of users. Using a Mann-Whitney statistical test again, we will be able to determine whether fake news broadcasters share more non-original content than other users. As this part of the work does not really require any preliminary processing of the data collected, the analysis can be found directly in the “Experimental results” section.

3.1.6. Method used to analyse user explicit features

Finally, we will achieve an analysis that focuses on the explicit features of the users, in an attempt to find characteristics that differentiate users who share fake news from those who share real news (**RQ4**). By “explicit features”, we refer to observable attributes directly available through user metadata on Twitter/X.

This final stage serves to contextualise the previous findings and evaluate whether there are any variables that influence the conclusions we have drawn regarding the presence of filter bubbles among users and their publication behaviour. Consequently, it will enable us to assess the robustness of the findings. To this end, we will draw on the work proposed by Shu et al. (2018). In their research, the authors divided their database into two groups: one group of users more likely to share fake news and one more likely to share real news. For each of these groups, comparative analyses for each feature are carried out. In line with our research and the data we have collected, the features we are going to study on are based on profile (whether the account is verified and when it was registered), content activity (number of posts and favourite actions taken), social network (number of followers and followings) and tweets (time period of activity, replies and retweets generated). Again, the characteristics we are going to analyse are already included in the data we have collected, so we do not really need to go through any processing steps. The statistical analyses of user features performed to check whether the differences between the two groups of users are statistically significant can therefore be found directly in the next section.

3.2. Experimental results

This part of the work presents the results of the analyses carried out on the data previously processed, in order to answer the various research questions.

Table 9 provides a summary of the hypotheses we have tested, with their statistical test and the p-value associated with each of these tests.

Hypothesis	Statistic test	p-value
Users who share fake news have a lower thematic diversity (RQ1).	Mann-Whitney U test	2.4433×10^{-7}
Users who share fake news tend to have views that are more marked, positively or negatively, and more distant	Mann-Whitney U test	0.17186
Users who share fake news tend to be more inclined to share recommended content (RQ3).	Mann-Whitney U test	2.8115×10^{-4}
Users posting fake news exhibit different behavioural or profile-based characteristics : • Registration time • Number of tweets • Number of tweets liked • Number of followers • Number of followings • Activity time • Number of replies on their tweets • Number of retweets on their tweets	Mann-Whitney U test	• 7.1698×10^{-01} • 1.8058×10^{-04} • 3.2299×10^{-05} • 6.2899×10^{-02} • 3.5357×10^{-02} • 9.3470×10^{-01} • 8.1999×10^{-02} • 3.6408×10^{-05}

Table 9: Summary of the hypotheses, their statistical test and associated p-value

3.2.1. Descriptive analysis of thematic diversity scores

We seek to test our hypothesis that users who share fake information have a lower thematic diversity (RQ1), potentially indicating that they are trapped in a filter bubble. For this purpose, a descriptive analysis was carried out on the thematic diversity scores obtained previously by separating the “real” and “fake” groups of users. Starting by applying a descriptive analysis helps us to identify trends visually and numerically before testing their statistical robustness with a formal test. The two distributions of diversity scores are visualised using a histogram. This histogram highlights any differences in shape, dispersion or centring between the diversity scores of the two groups of users. Descriptive statistics (mean, median, standard deviation, minimum and maximum) were also calculated to give a better understanding of the distributions.

Finally, a Mann-Whitney U test was applied to our data to statistically assess whether or not there is a significant difference between users who share fake news and those who do not. This non-parametric test was chosen because it does not rely on the assumption of normality of the data (Winters et al., 2010), which is appropriate for the non-Gaussian distribution of our diversity scores.

We now turn to the interpretation of the results of the descriptive analysis. We can see that the descriptive statistics of Rao’s diversity scores reveal differences between the two groups of users. Indeed, we see that users who shared real information have on average a higher diversity

score (≈ 0.74 versus ≈ 0.71 for the “fake” group). Another difference can be noted regarding the median. We notice a higher median among the users identified by the “real” label (around 0.77 while the median for the “fake” group of users is close to 0.74), indicating that their content is generally more diversified from a thematic point of view. In addition, both groups show some variability, but fake users show a slightly higher dispersion (higher standard deviation: $0.1356 > 0.1255$).

When we look at the histogram showing the distributions of thematic diversity scores on Figure 6, we see that both groups have a distribution concentrated between 0.6 and 0.9, but the “real” group has a more marked concentration around 0.8, while the “fake” group has a distribution tail more marked towards the low scores. On the basis of our dataset, we can therefore assume that a sub-group of “fake” users has a particularly low level of diversity, reinforcing the idea that their content is potentially confined to certain themes.

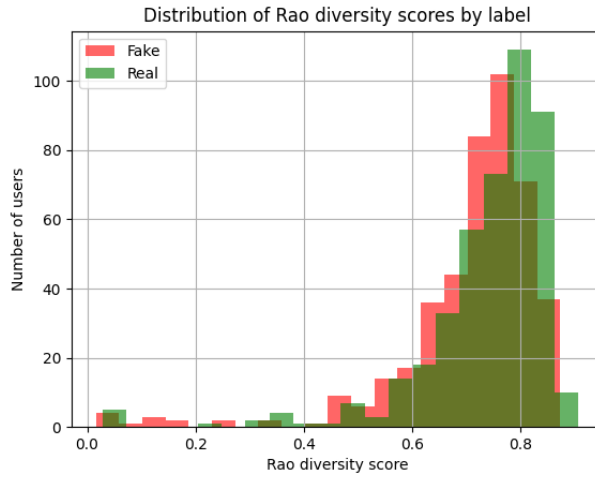


Figure 6: Distribution of users' Rao diversity scores by label

The graph shows that the diversity scores attributed to the two groups of users are distributed in the shape of a bell, despite a relatively stretched tail on the right-hand side. It is therefore relevant to check whether the data follow a normal distribution, in order to subsequently apply the appropriate statistical test to verify our initial hypothesis. As a first step, we carried out a visual diagnosis by forming the normal Q-Q plots associated with the two distributions (see Appendix 5). We can already see graphically that neither of the two distributions appears to follow a normal distribution. Indeed, the blue dots deviate overall from the red diagonal representing the theoretical normal distribution, especially at the extremities. To formally validate this intuition, we also performed a normality test. The scientific literature presents several tests for assessing the normality of the data. Mohd Razali & Yap (2011) have shown that of the most common normality tests, the Shapiro-Wilk test is the most powerful. This is why we decided to apply this test. Regarding the distribution of the diversity scores of users

who shared fake news, the p-value is equal to 1.0654×10^{-24} , which is well below the 0.05 threshold. This means that the null hypothesis, which assumes the normality of the data, is rejected. The same applies to the distribution of diversity scores associated with users who have shared real news (p-value = 4.7979×10^{-25}).

Since the normality assumption is not satisfied, we need to apply a non-parametric statistical test to verify our initial hypothesis. In our case, the null hypothesis assumes that there is no difference between the diversity scores associated with the “fake” and “real” groups. In contrast, the alternative hypothesis assumes that users who shared fake information have lower diversity scores than users who shared real information. The Mann-Whitney statistical test gives a very low p-value (2.4433×10^{-7}), confirming that the difference observed between the two groups is statistically significant, even if the average difference may appear moderate.

To conclude, the results statistically support our initial hypothesis, which states that users who share fake news are characterised by a lower diversity of content. This indicates that these users are potentially trapped in a filter bubble on Twitter/X, as the recommendations algorithm always provides them with similar content, generally tailored to their preferences and/or beliefs.

3.2.2. Descriptive analysis of sentiment scores

To test our second hypothesis (**RQ2**), we followed the same approach as for our first hypothesis concerning the thematic diversity of users. Firstly, we built the distribution of scores for the two groups of users in order to have a visual representation of the data. Next, we calculated the corresponding descriptive statistics so that we could analyse the trends from a numerical point of view. Finally, after checking the normality of the data, we performed a Mann-Whitney statistical test with the aim of confirming or refuting our observations with a more robust approach.

At first glance, the histogram of score distributions (Figure 7) shows that there does not appear to be any significant difference between the intensity of the opinions of users who share fake news and those who do not. Contrary to our hypothesis, it even seems that the distribution of scores for users associated with real news has a tail that is slightly skewed towards higher scores, with one outlier even having a score above 0.7, indicating more extreme sentiments. The descriptive statistics confirm our intuition. The means and medians of the two groups are very close, although slightly higher for the “real” group ($0.2488 > 0.2414$ for the mean and

0.2419 > 0.2388 for the median). The dispersion of scores is also slightly higher for “real” users (0.0639 > 0.0541).

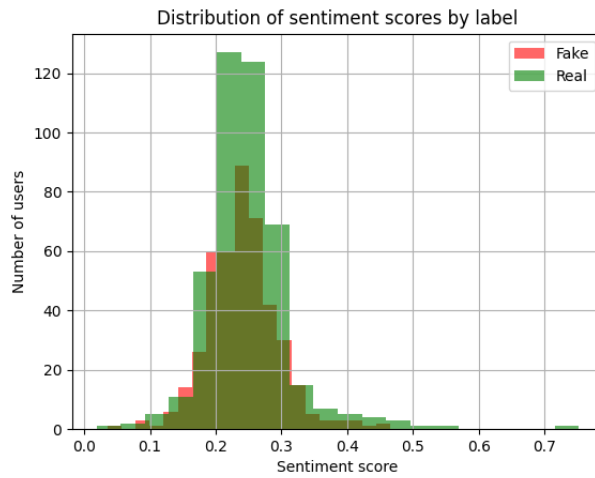


Figure 7: Distribution of users' sentiment scores by label

Once again, the graphical distribution of the sentiment scores suggests that the scores assigned to users in the two groups potentially follow a normal distribution. It is therefore necessary to verify this hypothesis, which is decisive for the choice of the statistical test to be used subsequently. From a visual point of view, the normal Q-Q plots (Appendix 6) associated with the two distributions show that the blue points generally follow the theoretical diagonal in the central zone. However, the points deviate from the red line at the extremities, especially on the right-hand side, which differs from a normal distribution. From a statistical point of view, the Shapiro-Wilk test formally rejects the normality assumption for both distributions, with p-values well below the 0.05 threshold (3.9072×10^{-7} for the “fake” group and 7.8609×10^{-19} for the “real” group). This indicates that the Mann-Whitney test is more appropriate for our further analysis.

Finally, the Mann-Whitney test definitively refutes our hypothesis. Indeed, our data does not allow us to assert that there is a significant difference between the two groups of users in terms of the intensity of their opinions, and more specifically that users who share fake news tend to have stronger opinions. The p-value of the statistical test is 0.17186, which is above the 0.05 threshold.

In conclusion, this analysis has enabled us to show that users in our data who share fake news do not necessarily have stronger opinions. Even though we have shown that these users are characterised by a lower diversity of content, this does not necessarily go hand in hand with greater emotional intensity.

3.2.3. Descriptive analysis of user publication behaviour

In order to deepen our understanding of the publication behaviour of users of our database, we also analysed the type of tweets they shared. Our hypothesis is as follows: “Users spreading fake news would be more inclined to share recommended (non-original) content, such as retweets or quoted tweets, rather than original tweets” (**RQ3**).

To test this hypothesis, we calculated for each user the proportion of non-original tweets (retweets and quoted tweets) among all the user’s publications we collected. A descriptive analysis was then carried out, followed by a Mann-Whitney statistical test to compare the distributions of these proportions between “fake” and “real” accounts.

If we interpret the descriptive statistics for the proportions of non-original tweets, we can see that, on average, users who share false information publish a higher proportion of retweets and/or quoted tweets (mean ≈ 0.46) than users who share real information (mean ≈ 0.37). This observation is even more marked for the median (median for the “fake” group ≈ 0.43 vs ≈ 0.24 for the “real” group). This means that at least 43% of the content shared by users in the “fake” group is non-original content for half of them, while half of the users in the “real” group have less than a quarter of their content that corresponds to retweets and/or quoted tweets. With regard to the standard deviations of the two distributions, we note that the two groups are characterised by a similar variability (0.3871 for the “fake” group of users and 0.3816 for the “real” group). These values indicate that there is some heterogeneity in behaviour within each group. This shows, for example, that not all users in the “fake” group systematically share non-original content, but the central trend remains that these users share a higher proportion of retweets and/or quoted tweets overall than “real” users.

According to the histogram (Figure 8), “fake” and “real” accounts are mainly distributed around the extremes (proportion of retweets and/or quoted tweets close to 0 or 1). On the right-hand side, we can see that a higher proportion of users in the “fake” group share a high proportion of non-original content (between 0.7 and 1), compared to users in the “real” group.

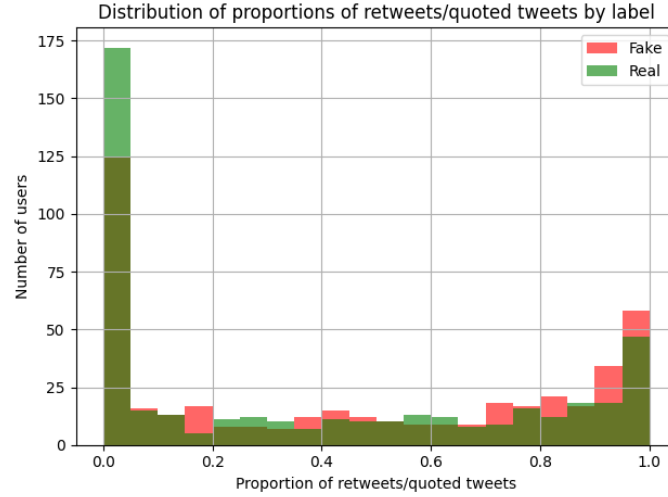


Figure 8: Distribution of users' proportions of retweets/quoted tweets by label

From the histogram, we can directly conclude that the data from the two groups do not follow a normal distribution. Consequently, the Mann-Whitney test is well suited to testing our hypothesis. From a statistical point of view, the null hypothesis assumes that there is no difference between the proportions of non-original tweets associated with the “fake” and “real” groups. On the other hand, the alternative hypothesis postulates that users who shared fake news are characterised by higher proportions of retweets or quoted tweets than users who shared “real” news. The Mann-Whitney test shows that there is a statistically significant difference between the two groups, with a p-value ($= 2.8115 \times 10^{-4}$) lower than the 0.05 threshold. “Fake” users therefore tend to share more non-original content than “real” users.

All these results confirm the hypothesis that users who share fake news are more inclined to share relayed content, such as retweets and quoted tweets, than users who share real news, who rely more on personal contributions. This may be explained by the fact that “fake” users are more sensitive to the content that is recommended to them, potentially influenced by Twitter/X’s algorithm.

3.2.4. Analysis of user features

Following all these analyses, we asked ourselves whether users disseminating fake news exhibit different behavioural or profile-based characteristics compared to those sharing real news (**RQ4**). The aim is to determine whether user-specific features explain the differences we have observed between the two groups of users so far, and therefore influence our interpretations.

To carry out this analysis, we drew inspiration from the work carried out by Shu et al. (2018). In their research, the authors compared the profiles of users sharing fake news with those

sharing real news. Their approach enabled them to identify whether or not there were significant differences between the two groups in order to better understand their behaviour and guide the detection of fake news. To this end, Shu et al. (2018) used t-tests to test for significant differences between the two groups, and analysed the distribution of values via the use of box plots or log-log cumulative rank curves.

Our analysis will focus on explicit characteristics obtained directly via profile metadata available on Twitter/X. These data were collected during the data collection phase by the second scraper when we collected user data. Four types of explicit features are studied to understand user behaviour.

As done in the work of Shu et al. (2018), the first type of feature is related to the profile of the user. To do this, we look at two attributes: account verification and registration date. Observing the distribution of verified accounts between the two groups of users could enable us to identify that users who share real news are more likely to have a verified account than users who share fake news. The registration date, meanwhile, lets us know whether accounts associated with the “fake” label in our database tend to be older, for example.

The second type of feature relates to the content. The attributes considered are the number of posts and the number of favourite tweets of the users. Analysis of these attributes could enable us to see that “fake” users generally publish more than “real” users, and that they also bookmark more often other users’ tweets.

The social network aspect is the third type of feature explored. We use the number of followers and followings for this part of the analysis. These attributes could allow us to demonstrate that “fake” users have fewer subscribers on average than “real” ones, or that they follow more accounts. A TFF ratio calculation ($TFF = \frac{\#Follower+1}{\#Following+1}$) is also applied to this data, enabling us to see if one type of user is more popular than another.

Finally, the last type of feature to be analysed is based on tweets. This kind of characteristic has not been applied by the authors Shu et al. (2018) as they only focus on user data. Analysing the data on users’ tweets makes sense in the context of our research, given that we have this data. We may be able to discover a specific trend regarding the tweets of users in the “fake” group that does not exist in the tweets of users in the “real” group. We observe whether there is an impact in terms of their activity time (how long it is between their oldest tweet and their most recent tweet among the users’ tweets we collected), the number of replies received on their tweets, and the number of times their tweets have been retweeted by other users.

Let's now look at the results of these different analyses to see whether or not there are any significant differences between our two analysed groups.

Various approaches were used to analyse these results. First, box plots and cumulative log-log rank curves were used to perform visual and statistical analyses on the examined characteristics. Then, we performed a Shapiro-Wilk normality test. This test enabled us to determine which test is more appropriate between a t-test (if the distributions are normally distributed) and a Mann-Whitney test (if the distributions do not follow a normal distribution), to assess whether there is a significant difference between the two groups of users for each studied attribute. Appendices 7 to 14 show that our data never follow a normal distribution. This is why Mann-Whitney tests were systematically applied to our data.

With regard to the first profile attribute, we note a low proportion of verified users overall (less than 4% of users in each group). This indicates that the vast majority of users, whether sharing fake news or not, are not verified. The graph provided (Appendix 15) confirms the similar proportions of verified accounts between the two groups of users, allowing us to assert that a verified user is not a distinguishing criterion between the two groups. Within our sample, we can therefore affirm that there is no strong link between being verified and the veracity of the content shared (fake vs real news).

For user account registration dates, the box plot (Appendix 16) represents the distributions of the number of days elapsed since user registration. The median of the “fake” group is close to the median of the “real” group, indicating that the accounts were created around 13 years ago. Our sample therefore includes rather old accounts, whether they share fake news or not. We can also add that the distributions are largely overlapping. Furthermore, the p-value confirms that there is no significant difference between the two groups of users ($0.717 > 0.05$). The date of creation of the Twitter/X account is therefore not a clear discriminating factor between the two groups in our sample.

If we now look at the content-related features and, more precisely, at the total number of tweets posted by each user, we obtain a p-value close to 1.81×10^{-4} , which is well below the 0.05 threshold. Hence there is a statistically significant difference between the two groups. There is also a difference in the medians for the number of tweets within the groups. The median of the “fake” group is equal to 19,586, while the median of the “real” group is 11,419. The log-log plot (Figure 9) used for this attribute represents the inverse cumulative distributions of the number of tweets. The X axis is logarithmic and shows the number of tweets, while the Y axis

indicates how many users have posted at least x tweets. In this graph, we can see that the “fake” and “real” curves show a clear separation: “fake” users tend to have published more tweets than “real” users over the whole distribution. Indeed, we can see that up to a certain threshold (around 10^5 tweets), the “fake” curve is positioned above the “real” one. But from a very high threshold (above 10^5 - 10^6 tweets), the “real” curve rises above the “fake” one, meaning that more “real” accounts post a very high number of tweets than “fake” accounts. In view of this change, we wanted to find out more about the users who have posted a number of tweets greater than 10^6 . Looking at our data, we saw that only one “real” user has a number of tweets well in excess of 10^6 . This is the ‘AmazonHelp’ account, which has 12,817,192 tweets. This Twitter/X account is Amazon’s official support service, dedicated to answering questions and resolving customer problems in several languages, which explains why this user has a very large number of posts published. The graph therefore shows that “fake” users are on average more active (confirming the median analysis and the statistical test), although a “real” user is associated with the highest number of tweets published in our sample. We can therefore say that “fake” users seem to publish significantly more tweets than “real” users, which allows us to state that, based on our sample, “fake” accounts tend to publish more posts than “real” accounts.

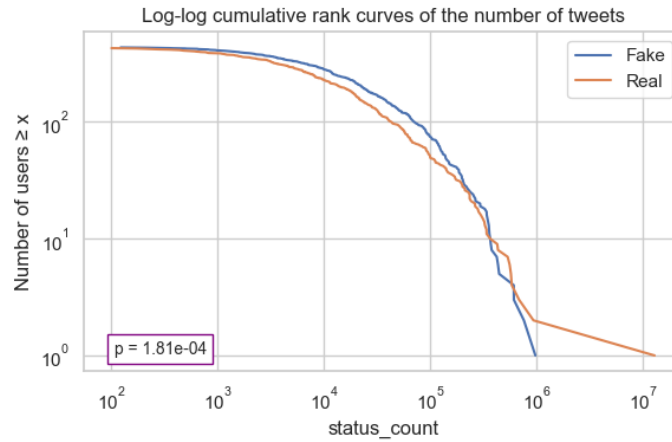


Figure 9: Log-log cumulative rank curves of the number of tweets posted

The total number of tweets liked per user is a discriminating feature. The “fake” group has a higher mean value equal to 32,070.39, compared with a mean value of 21,755.5 for the “real” group. We also see a higher standard deviation value for the group of “fake” users ($79,402.58 > 68,999.68$). Analysing the results obtained (as we can see in Figure 10), we note that the curve representing “fake” accounts is systematically above the curve for “real” accounts. Users who share fake news in our sample are therefore more likely to like more content than the others. A difference in median is also observed for users in the “fake” group, with a value almost four times higher than that associated with users in the “real” group ($3,939 > 1,063$). This finding is

confirmed by the p-value of 3.23×10^{-5} obtained by the Mann-Whitney test, which indicates a statistically significant difference. These results therefore show that “fake” users seem to like more content than “real” users.

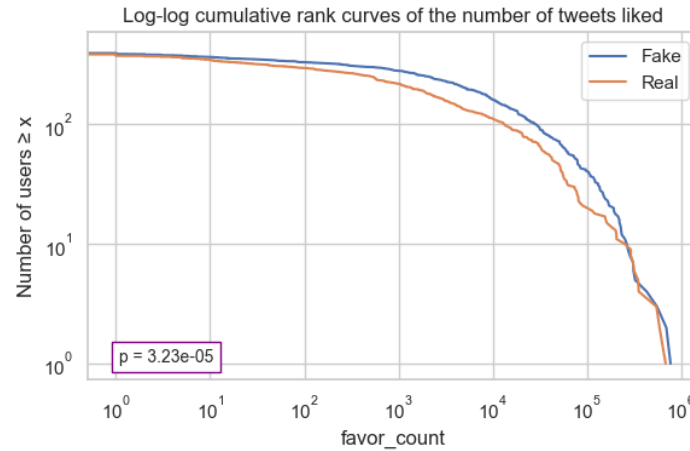


Figure 10: Log-log cumulative rank curves of the number of tweets liked

Now let’s take a look at the network-related attributes. The box plot (Appendix 17) showing the number of users’ followers reveals very similar distributions between the two groups, but with a slight tendency for “fake” users to have more followers. The median for the “fake” group is 529 and that for the “real” group is 422. The box plots also show a wide dispersion within the two groups. Although the median for “fake” accounts is slightly higher, this difference is not statistically significant. Thus we cannot clearly distinguish users who share fake news from users who share real news in our sample, based on their number of followers.

However, a different observation can be made for the number of followings. Despite similar distributions and almost identical medians (492 accounts followed by “fake” users and 445 accounts followed by “real” users), Figure 11 shows that “fake” users tend to follow more accounts than “real” users. Furthermore, we can confirm that there is a significant difference within our database, as the Mann-Whitney test returns a p-value equivalent to 0.035357. In view of these results, we can conclude that the number of accounts followed by a user can be a useful behavioural indicator regarding the sharing of fake news.

The TFF ratio is used in our research as an indicator of influence or reputation. A high ratio means that a user has more people following him/her than he/she follows people. Based on our data, we can see on Figure 12 that “real” users have a higher average TFF ($2,263.64 > 1,534.66$), suggesting greater perceived credibility or a wider audience. The visualisation of the ratio shows an asymmetrical distribution and an overlap of the two histograms, suggesting that the

differences between the groups are subtle. We can also note that the majority of both groups have a ratio close to or less than 1.

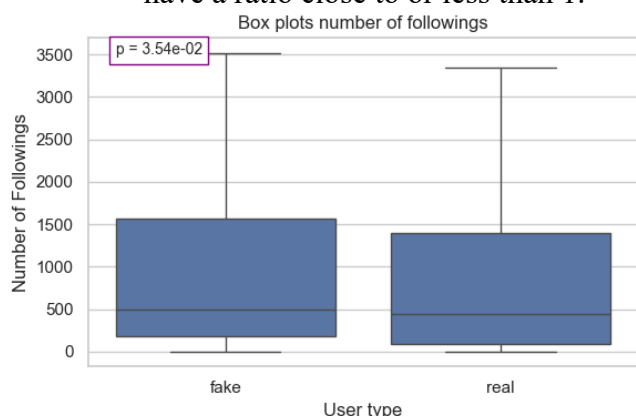


Figure 11: Box plots of the number of followings

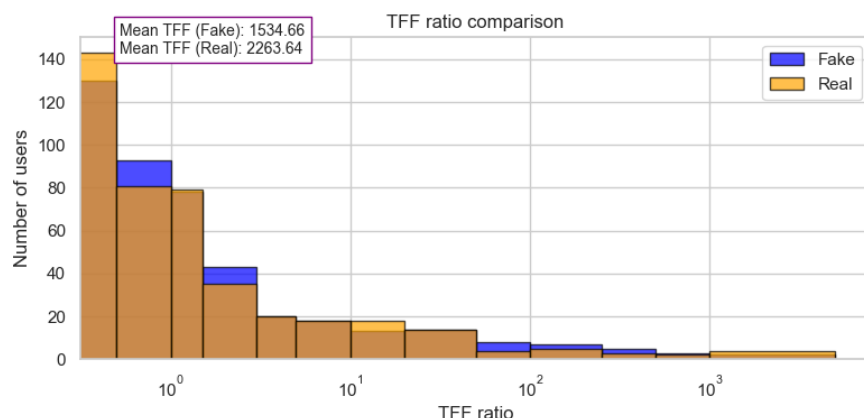


Figure 12: Comparison of TFF ratio distributions by label

Let's now analyse the time window of the tweets. When we look at the temporal extent of user activity on Twitter/X (time interval between the oldest and the most recent tweet), we obtain very close medians between the groups (226 days for the “fake” group and 235 days for the “real” one). The Mann-Whitney test we applied to these data gave a p-value of 0.935, indicating that there was no significant difference between the two groups. The box plots in Appendix 18 and the cumulative curves in Appendix 19 show extremely similar distributions. At all levels of activity (short, medium or long duration), the groups show equivalent profiles. This characteristic therefore does not allow us to differentiate between the two types of users.

To conclude this behavioural analysis, we look at the engagement generated by users, i.e. the average number of replies and retweets per tweet. The average number of replies per tweet is very low (less than one reply per tweet). Although “fake” users have a slightly higher median, the p-value of 0.082 indicates that there is no significant difference with the group of “real” users. The boxplot (Appendix 20) and cumulative curves (Appendix 21) confirm this by showing overlapping distributions. “Fake” accounts appear to receive slightly more replies per tweet than “real” accounts, but this difference is not significant. This result suggests that “fake” accounts can sometimes generate interactions, but not systematically enough to be differentiated from “real” accounts on this criterion.

On the other hand, accounts disseminating fake news seem to generate more retweets, indicating a greater potential for virality. The p-value of 3.64×10^{-5} indicates a significant difference proving that “fake” users receive on average more retweets than “real” users. This finding is supported by the huge difference between the medians. Indeed, the median obtained for the “real” group (38.1 retweets/tweet) is much lower than the median obtained for the “fake” group (418.8 retweets/tweet). The box plots (Figure 13) support this difference and show a

greater dispersion of retweets for users belonging to the “fake” group. The cumulative log-log curve (Figure 14) shows that, for almost all retweet values, more “fake” users reach high levels. We can therefore say that “fake” users generate significantly more retweets per tweet than “real” users, which is a good indicator for differentiating the label associated to each account. This analysis of features reveals that many of the attributes studied (related to profile, activity, network and tweets) clearly differentiate the two groups. In fact, several elements proved to be significantly discriminating: the number of tweets posted, the number of likes given, the number of followings and the number of retweets per tweet. These results lead us to say that, based on our sample, users sharing fake news tend to publish more tweets, to like more content and to obtain a greater number of retweets under their tweeted posts – suggesting a higher potential for virality. They also tend to follow more people on Twitter/X than users from the “real” group. These results enrich our study while highlighting specific behaviours associated with misinformation.

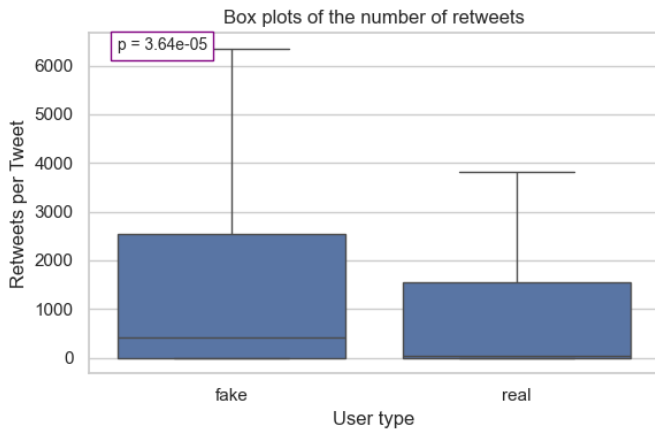


Figure 13: Box plots of the number of retweets

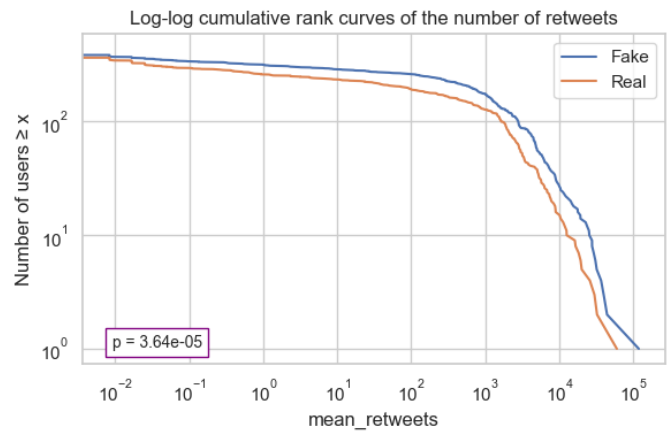


Figure 14: Log-log cumulative rank curves of the number of retweets

One limitation can be raised following this analysis. We ourselves selected the features to be analysed on the basis of the data we were able to collect. However, there are potentially other relevant attributes that could be analysed to differentiate users who share fake news from those who do not, such as implicit user features that are more difficult to identify.

4. Discussion

With this study, we sought to determine whether a link could be established between the sharing of fake news on social networks and the filter bubble phenomenon. In addition, we also wanted to take a closer look at the behaviour of people who share fake news online.

Firstly, we wanted to evaluate whether users who share fake news are more likely to be trapped in a filter bubble by studying the diversity of their content (**RQ1**). The aim of this first analysis was to verify our hypothesis that users who share fake news are characterised by a lower thematic diversity of content than users who share real news. The results we obtained confirm this hypothesis, given that users who have shared false information have significantly lower thematic diversity scores than others. This finding leads us to believe that those users are more likely to be trapped in a filter bubble on the social network because their recommendation algorithm tend to provide them with less diverse content, potentially similar to their preferences, ideas and beliefs.

Even if the results show that there does indeed seem to be a relationship between enclosure in a filter bubble and the sharing of fake news, the direction of causality between these two phenomena is difficult to determine. On the one hand, it is the enclosure in a filter bubble that could lead users to share fake news. As Bakshy et al. (2015) have shown, algorithmic personalisation reinforce user bias through repeated exposure to similar information. As a result, users' ability to identify fake news could be altered, leading them to become more credulous about this type of content and to share it more easily. In this case, the reinforcement of cognitive biases, altering the user's judgment, is a mediating variable that explains the causal relationship. On the other hand, the fact that users share false information could reinforce their informational confinement. Indeed, Pariser's (2011) theory of algorithmic personalisation warns that users are increasingly exposed to homogenous content aligned with their past behaviour. Fake news often has specific presentation characteristics (Wogalter & Mayhorn, 2008) and is generally sensationalist content that generates engagement. The algorithm could thus tend to react more strongly to this type of content shared by the user and recommend even more similar content to him/her, thus reinforcing the user's bubble. The algorithmic engagement generated by the sharing of fake news therefore explains this causal link.

It should be noted that certain variables can also influence the relationship between filter bubbles and the sharing of fake news. Among the moderating variables we can identify the level of digital literacy and the age of the user. People who are more comfortable using digital technologies and have a deeper understanding of how they work will tend to be potentially less

influenced by algorithmic bubbles. Complementarily, people from Generation Z have grown up with the Internet and will therefore be less likely overall to be fooled by filter bubbles, while children and older people might tend to be more easily influenced and therefore be more likely to share false content. Finally, the user's level of social interaction may also have an impact. Users who are highly engaged in online communities may already be more inclined to face informational confinement, which already amplifies the filter bubble phenomenon beyond the effect that sharing fake news can have on algorithmic engagement.

All these assumptions could be tested empirically. For mediating variables, mediation analyses should be carried out, for example using the Baron and Kenny method or the Sobel test to verify the significance of the mediating effect. For the validation of moderating variables, interaction analyses can be conducted.

Secondly, we looked at filter bubbles from the perspective of the emotional intensity of opinions, by testing the hypothesis that users who share fake news on social media tend to have stronger opinions (**RQ2**). However, the results obtained following this second descriptive analysis did not allow us to affirm that there is indeed a link between the intensity of opinions and the veracity of the news shared by the user. The Mann-Whitney test refuted our hypothesis, showing no significant difference between the sentiment scores of the two groups of users. Consequently, “fake” users do not necessarily have stronger opinions. Even if these users are enclosed in a filter bubble in terms of the diversity of content they are confronted with, we have been unable to show that their opinions tend to be more extreme.

However, this does not prevent users who share fake news from having reinforced opinions. Andrea (2022), Karlsen et al. (2017), Pariser (2011), Rowland (2011), and Sunstein (2001) argued that repeated exposure to homogeneous content leads to belief reinforcement and polarisation. As demonstrated through RQ1, “fake” users are more likely to encounter less diverse content, which can therefore lead to them being more polarised in their opinions. Our findings may therefore simply suggest that polarisation does not necessarily manifest itself through visible emotional intensity. Users may be ideologically polarised without resorting to emotionally extreme language or they may deliberately moderate their tone to appear credible or avoid confrontation on the platform.

We then looked at the type of content that users publish. We sought to show that users who share fake news tend to share more relayed content (retweets and quoted tweets) than original tweets (**RQ3**). A significant difference was demonstrated by the Mann-Whitney test,

confirming our hypothesis. This finding leads us to the following assumption: since they are confronted with more restricted content, potentially aligned with their preferences, users who share false information are more sensitive to this recommended content. As a result, they are more likely to re-share this content via retweets and quoted tweets, than to post original, self-initiated tweets. This may reflect a more passive behaviour, reactive to algorithmic flows, while “real” users seem more proactive in publishing original content. Such passivity can facilitate the viral spread of false information, often shared without prior verification. We can notice that it is consistent with the “passive sharing” behaviour described by Pennycook & Rand (2019), where content is circulated impulsively without deep reflection. Furthermore, it supports Cinelli et al. (2021)’s insights that algorithmically recommended content is more likely to be shared by users exposed to less diversity. The higher reliance on non-original tweets may reflect an increased influence of recommender systems, especially those based on collaborative filtering, which amplify popular or emotionally salient posts, potentially explaining the higher circulation of misleading content (Vosoughi et al., 2018).

Nevertheless, it is important to keep in mind that certain factors may play a moderating role in the relationship highlighted above. For example, a user’s general publishing behaviour (such as his/her overall volume of shared tweets) could influence the observed pattern. Very active users may naturally produce a greater number of retweets and quoted tweets, regardless of the veracity of the content they share. In this case, the difference between the two groups of users in terms of retweeting behaviour may be less pronounced, suggesting that the activity levels of the users under study may attenuate or amplify the differences observed between the two groups.

Finally, we carried out a final analysis based on different user characteristics in order to detect any differentiating features between users who share fake news on social media and those who do not (**RQ4**). Such distinctive attributes would enable us to better understand the profile of users who share false information, and potentially explain the results highlighted so far. Through our analyses, we found that “fake” accounts tend to publish more tweets and to like more content on Twitter/X. These users also have a propensity to follow more people on the platform than “real” users. Moreover, their tweets tend to generate more retweets within their community. This increased social activity can boost their visibility in other people’s news feeds. As a result, false information can circulate more widely thanks to algorithmic engagement loops, where liked or retweeted content is amplified regardless of its veracity. These behavioural features are in line with the idea of algorithmic feedback loops (Narayanan, 2023):

increased user engagement (likes, follows, shares) encourages the algorithm to offer similar content, thereby reinforcing the filter bubble over time.

Nevertheless, it is important to acknowledge a number of limitations that may have influenced the outcomes of our research.

Firstly, the data on which we are focusing our analyses comes solely from the Twitter/X social network. Secondly, we are analysing a limited sample of users/tweets due to restricted resources, which may not be representative of reality. In addition, “manual” filtering had to be applied to our data in an attempt to retain only as many users relevant to our research as possible. However, some users deemed irrelevant to our work could not be filtered out at this stage. Moreover, one important limitation lies in the distinction between a news feed with little diversity and personal consumption with little diversity. Our analysis evaluates thematic diversity based on the tweets users post, retweet or quote. However, a low-diversity timeline does not necessarily mean that the user is not exposed to diverse content overall. It is possible that some users actively seek out divergent opinions or explore varied topics privately (e.g., via searches, passive reading, etc.), without necessarily sharing these contents. Therefore, what we observe may reflect visible expression more than personal exploration. Our methodological approach, focusing on publicly shared content, does not allow us to capture such exploratory or non-public consumption behaviour, which constitutes a significant limitation in identifying the full extent of the filter bubble effect. Finally, we separate users into two groups on the basis of whether or not they shared at least one fake news item according to the website Gossip Cop. This definition of “fake” and “real” users does not fully represent the real complexity that exists on social networks: users can share both real news and fake news. Consequently, the stated hypotheses may simplify a more nuanced reality.

On the basis of these various limitations, we can identify various tracks for future research. One idea would be to extend our analyses to other platforms such as TikTok, for example. Su (2023) states that the large number of users on TikTok leads to the rapid spread of fake news. This assertion confirms that extending our analysis to this platform makes sense. Another possibility of research would be to carry out the analysis on a larger sample of data that could provide a more concrete representation of reality. Finally, integrating clustering methods to identify typical user profiles within users who share fake news and those who share real news could be a potential area for future research. Also, we are focusing on explicit behavioural dimensions, but looking at implicit behavioural dimensions or psychological characteristics could be relevant for further research.

To conclude, our results as a whole show a consistent behavioural profile among users who share fake news. These users are characterised by greater activity on the platform (high number of posted tweets and likes), a tendency to follow many accounts, and a marked tendency to reshare recommended content (retweets and/or quoted tweets). These elements suggest a form of intense interaction with the Twitter/X algorithmic system, which could play a key role in reducing the diversity of the content faced by those users. Consequently, these variables could explain the fact that users who share false information seem to be more confined in filter bubbles. It is therefore not only what users post that matters, but also how they interact with the platform itself. In other words, it is not only the algorithm that passively traps these users in an information bubble, but a feedback loop between their behaviour and the recommendations they receive.

This complex self-reinforcing dynamic between interaction with the algorithm, enclosure in a filter bubble and the sharing of fake news ultimately shows us that the fight against fake news cannot be limited to fact-checking. It also requires more refined regulation of recommendation mechanisms and greater accountability on the part of platforms. It is therefore essential to consider the implications of these findings in terms of algorithmic design, platform governance and public information policies.

5. Managerial implications

In the light of the results we have obtained, it is now relevant to analyse their concrete managerial implications, in order to identify possible levers for action to limit the spread of false information and mitigate the effects of filter bubbles.

Our analyses have shown that users who have shared fake news are exposed to content that is less diversified and consequently, are more likely to be stuck in a filter bubble. As a result, managers of platforms such as social networks should try to incorporate more explicit diversification mechanisms into their algorithms to avoid informational lock-in. Indeed, it is important to seek to improve diversity given that focusing solely on relevance can be detrimental to the discovery of new content or long-term user satisfaction, reinforcing filter bubbles and self-confirmation (McNee et al., 2006; Nguyen et al., 2014; Ziegler et al., 2005). A great deal of work in this area already exists. Among this work, researchers have explored a number of different approaches: simulation and reranking (L. Zhang et al., 2025), integrating fairness with diversity (Zhao et al., 2025), taking into account individual preferences for diversity and serendipity (Yin & Zhao, 2024), etc.

In addition, our analysis of user features enabled us to identify specific attributes that distinguish users who share fake news from those who share real information. This shows that it is possible to identify to a certain extent profiles at risk of spreading fake news. The people responsible for combating misinformation could develop preventive detection models based on the distinctive characteristics of profiles that are potentially more vulnerable to the spread of false information. In addition to our work, other studies have sought to identify profiles at risk of propagating fake news on the basis of behavioural, social, linguistic or psychological characteristics (Grinberg et al., 2019; A. Guess et al., 2019; Shu et al., 2018).

Another element that caught our attention when writing this thesis is that users are not always aware that they may be facing restricted content, sometimes manipulated by the algorithm. For this reason, we believe it is crucial to make users aware of their informational exposure in the current societal context, where we feel ironically overloaded with a variety of information. It would be the responsibility of platform managers (especially communication managers) to set up awareness campaigns to help users understand how their actions influence what they see online. For example, information messages on social networks aimed at educating users on the subject could be an idea worth developing.

Finally, still with the aim of limiting the filter bubble phenomenon on social networks, integrating diversity metrics into performance indicators is a potential future solution to be implemented by these large companies. Platforms often measure algorithmic performance in terms of engagement (clicks, likes, page views, etc.) (Bozdag, 2013; Helberger et al., 2018) but rarely in terms of diversity. Thus, including an informational diversity indicator as a Key Performance Indicator (KPI) could help balance performance and quality of user experience.

6. Conclusion

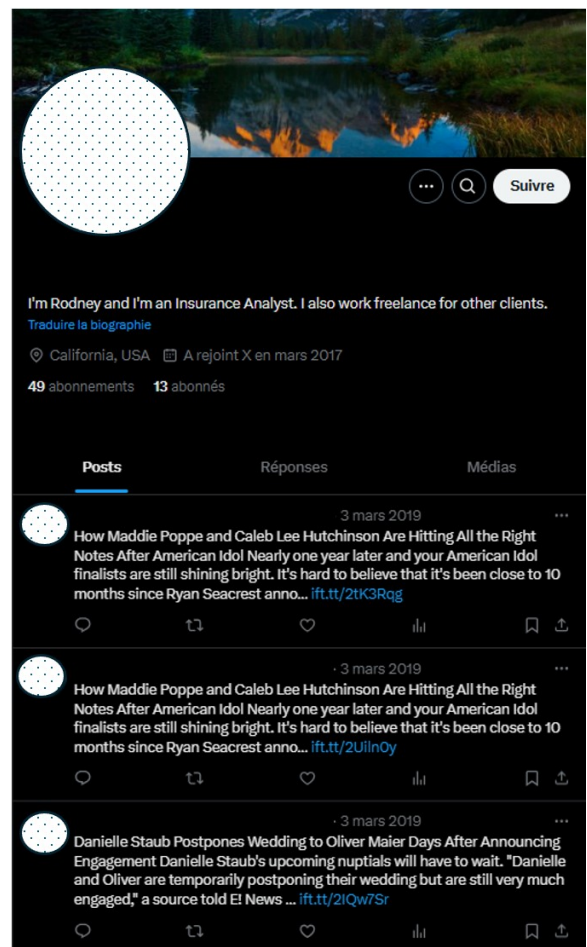
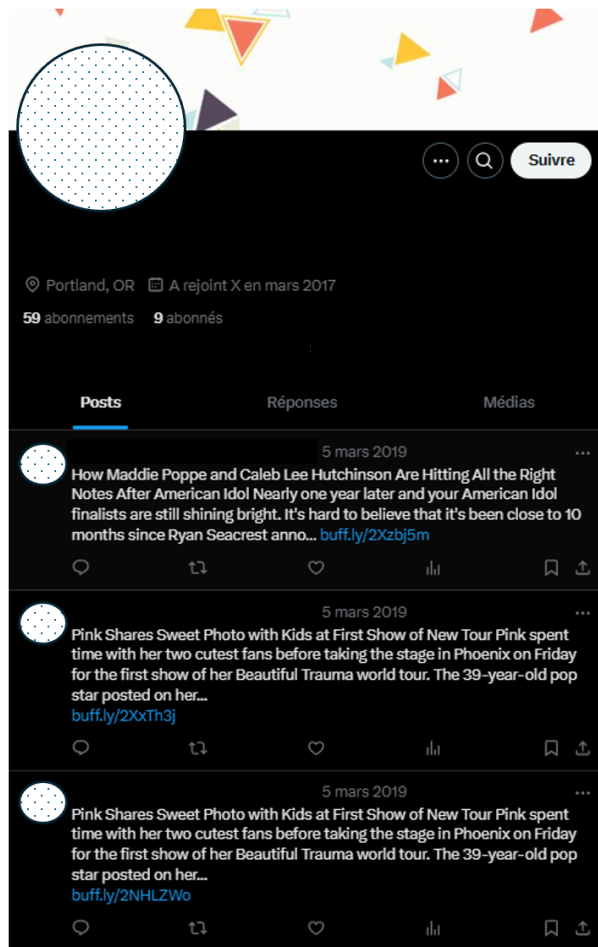
In a context where social networks are a growing part of our relationship with information, our study aimed to analyse in more detail the relationship between the sharing of fake news and enclosure in a filter bubble on social media. Furthermore, we also sought to learn more about the behaviours that characterise users who share fake news online.

To this end, we adopted a quantitative approach based on descriptive analyses, combining natural language processing methods, topic modelling, sentiment analysis and user profile analysis. These different approaches have been applied to data that we collected ourselves. The data collection phase is based on news articles verified by the website Gossip Cop. After various stages of scraping and filtering, we analysed the tweets of 864 users divided into two groups: 429 users who had shared at least one piece of real news and 435 users who had shared at least one piece of fake news according to Gossip Cop.

The various results we obtained from our analyses confirmed several of our hypotheses. We were able to demonstrate that within our data, users who share fake news are characterised by a significantly lower thematic diversity. This result therefore supports the idea that these users are more likely to be trapped in a filter bubble. In addition, we observed that “fake” users tend to share more non-original content (in the form of retweets and quoted tweets), indicating potentially a greater sensitivity to the content recommended to them. We also identified that users who share fake news on Twitter/X are more likely to post, like content and generate more retweets. They also tend to follow more accounts on the platform than users who share real news. However, no significant difference was found in the emotional intensity of opinions between the two groups of users. Hence lower thematic diversity is not necessarily accompanied by stronger opinions. Without claiming to establish strict causalities, these results consequently suggest a link between low informational diversity, enclosure in a filter bubble, user behaviour and the sharing of fake news. These findings underline the need for further reflection on the role played by digital platforms in structuring information. They also remind us of the importance of promoting greater algorithmic transparency and strengthening users’ critical skills with regard to online information.

7. Appendices

Appendix 1: Examples of spam users on Twitter/X



Appendix 2: User with a limited number of posts in his/her timeline

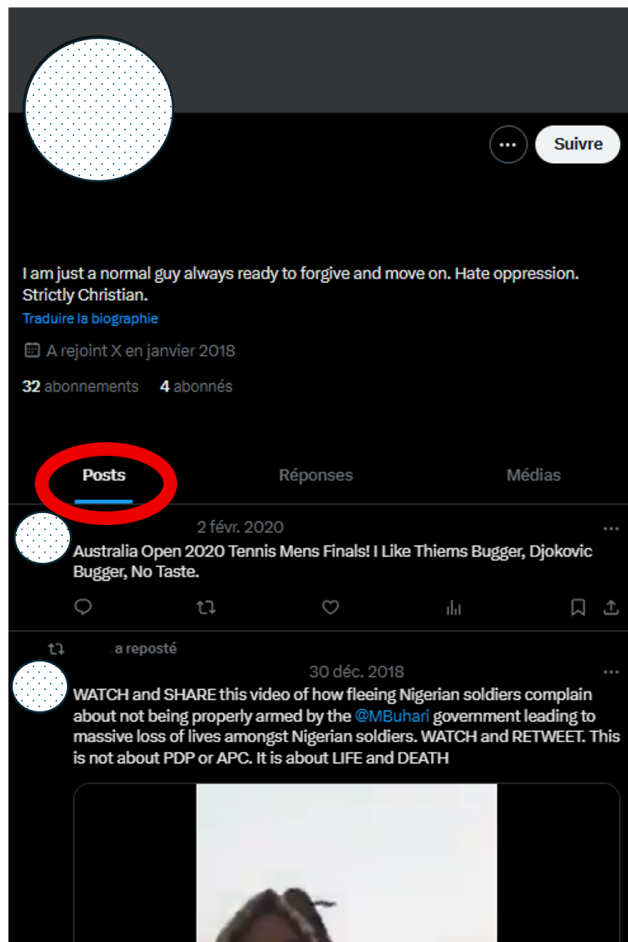
Profile Summary:

- 7 105 posts
- 492 abonnements
- 3 120 abonnés
- Da Tr 🍷th
- A rejoint X en juillet 2009

Timeline Posts:

- 1** a reposté 21 avr. I love fresh starts, you can set the tone, do it differently & do it better. 13 replies, 6 k retweets, 10 k likes, 440 k views.
- 2** a reposté 16 mars 3 G's .. Grinding , God & Growth. 158 replies, 6 k retweets, 16 k likes, 530 k views.
- 3** a reposté 11 mars Soooo ready for paradise 🍷 the world was such a beautiful place because she was in it. Feels empty , like a big chunk is missing 🍷 oh mama 😭 1 reply, 1 like, 55 views.
- 4** a reposté 11 mars I miss calling her everyday, talking to her, seeing her, aggravating her. She was the biggest part of me . 1 reply, 1 like, 3 views.
- 5** a reposté 11 mars The BEST mother you could of ever asked for. Oh how i miss my sweet mama 1 reply, 2 likes, 100 views.
- 6** a reposté 12 févr. This isnt goodbye mama but see you soon 🍷🍷 you taught us everything we needed to continue going strong in this old system. We promise to make you proud & stick together 🍷🍷 see you soon! 🍷🍷 John 5:29 Jehovah promise is so real 🍷 1 reply, 2 likes, 59 views.
- 7** a reposté 12 févr. I miss my mama so much 🍷 1 reply, 2 likes, 3 views.

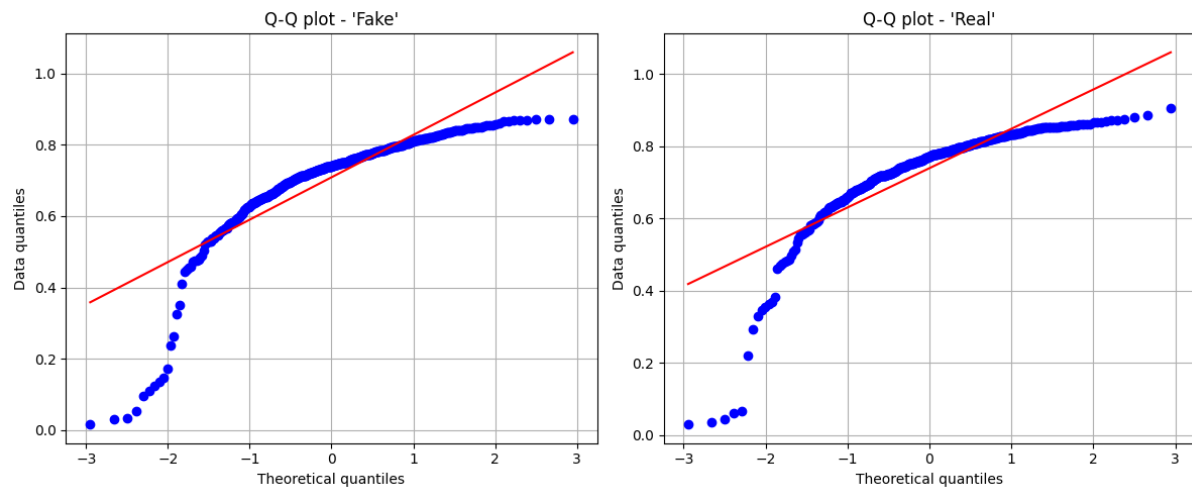
Appendix 3: Example of a user's Twitter/X timeline



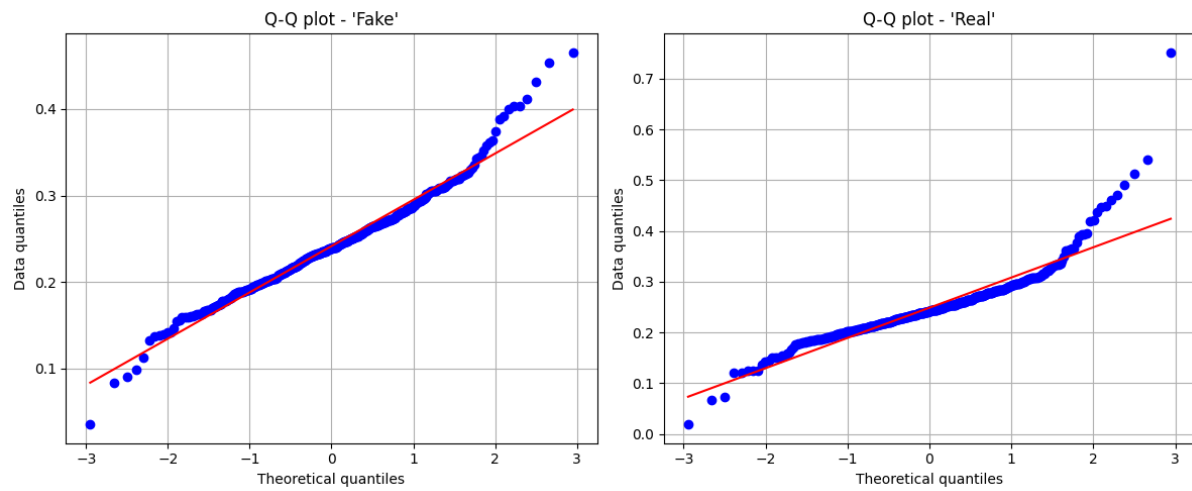
Appendix 4: Extract from the output of the MALLET command providing the most representative words for each topic

```
0 0,08885 work make need use health million new year help amp business company billion high tax family pay plan support covid
1 0,06171 love life god thank day family happy good amp today make world give help heart one person support may child
2 0,08072 toronto check story today ryan real jeff photo gta download estate mississauga buy insurance mplusplace alex call house give free
3 0,03267 israel attack person afghanistan gaza country world kill terrorist say child starmer palestinian news minister call support hama israeli government
4 0,01461 food eat recipe tiktok video check second delicious yum taco lady chicken make annieklein bettertogether like glamb easy cake dinner
5 0,02131 win chance amp ticket get follow enter winner see p.m give sale giveaway end like day competition show prize retweet
6 0,04499 star movie season new film trailer play show series watch actor episode next celebritynew imdb first official netflix oscar john
7 0,01597 die rock art late thanks day bear guitar music buy original fat song edm good wiba today band amp abstract
8 0,01642 storm weather snow high rain via sky space moon hurricane wind low alert good amp earth today nasa morning flakethemixtapemogul
9 0,01562 ukraine russia russian war military putin attack ukrainian force drone trump alaska president strike operation missile today world nuclear peace
10 0,02491 live listen watch tonight new p.m get show stream day episode join app amp music tune miss podcast channel part
11 0,04534 game win team first play player sport final football season league fan champion goal time versus nba cup back coach
12 0,03655 via new baby justin news celebrity amp post bieber photo kim first lady kardashian perry dating harry wedding show love
13 0,00471 travel space daily web nft late max wordpress grok slack xbox fintech metaverse cybersecurity xai quest virtualreality muvc instagram devop
14 0,01089 weight loss share surgery friend plastic news blog life daily nintendoswitch wise iam bjp style award india amp mchammer raise
15 0,00496 thousand home realestate comp flipthishouse wholesalehome realestateinvesting fixerupper fliporflip wholesalehouse rrei mls renovate good level warhammer brick
warhammercommunity amp forest
16 0,02084 check good makeup fashion look style dress beauty hair add ebay list shoe size closet age gray poshmarkapp poshmark woman
17 0,03466 trump via news say court case judge break new charge president arrest justice claim biden trial prison former report crime
18 0,07142 trump president vote american person state amp say biden republican election america democrat country donald get would bill right million
19 0,01218 year cannabis car marijuana today ago may auto onthisday daysofmotoring race classiccar team first versus nascar learn begin formula prosperity
20 0,03632 music new album song amp love release kid listen taylor video single tour target drake year artist award top million
21 0,01944 book bond film new vegas jame las read art story cinema amp jamesbond via series artist take feature museum home
22 0,00958 one person today via follow follower check unfollower stat automatically new unfollowed despacito road ndavenue track find ling resale xyz
23 0,24821 get like million person know say one make good time see want think would look thing need love year come
24 0,03288 free get please game new order link amazon use help million join iphone need earn day launch good check see
25 0,02481 post social media photo new instagram twitter news facebook follow video content blog marketing get via free share use check
26 0,06869 day happy good year new today get great weekend morning birthday time christmas week come night see last may celebrate
27 0,03971 police year shoot man old school kill person officer county arrest woman shot two city one dead say child fire
28 0,01138 dog please animal help cat need pet rescue via foxnew secio android app food donate adopt old puerto support share
29 0,02985 youtube video like via new amp movie playlist add music official review episode time york miley cyrus nyt love audio
```

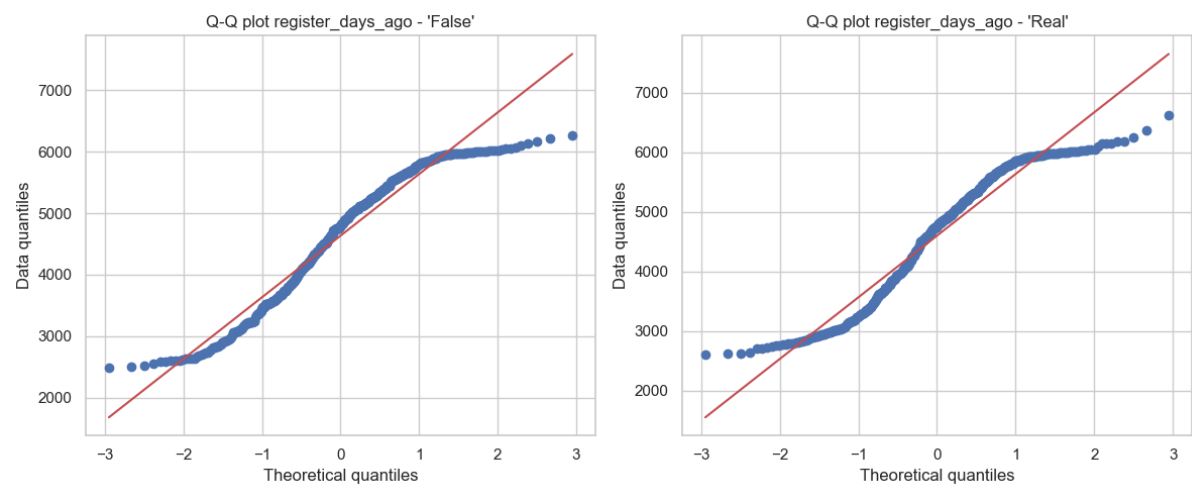
Appendix 5: Normal quantile-quantile plots of users' Rao diversity scores by label



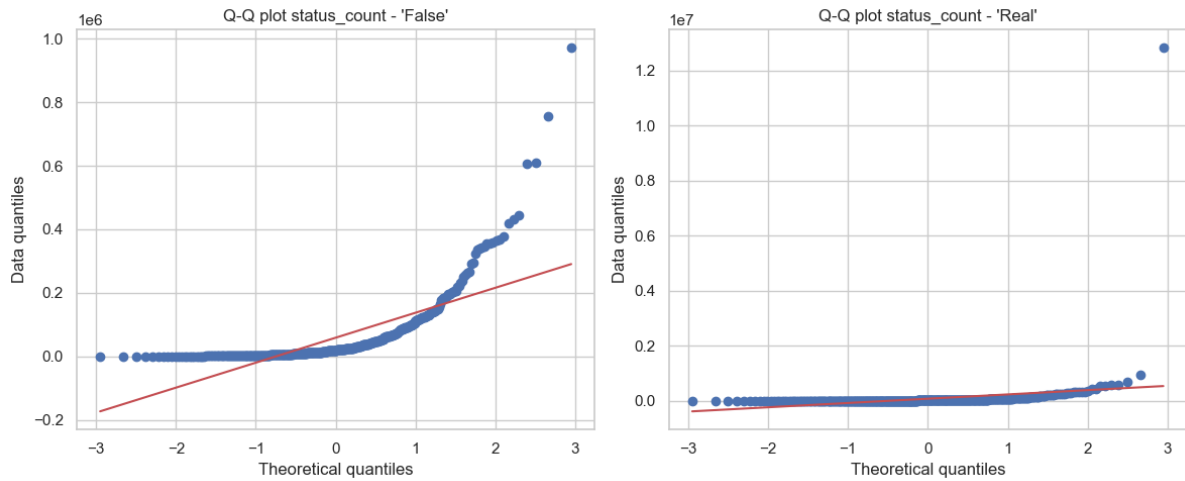
Appendix 6: Normal quantile-quantile plots of users' sentiment scores by label



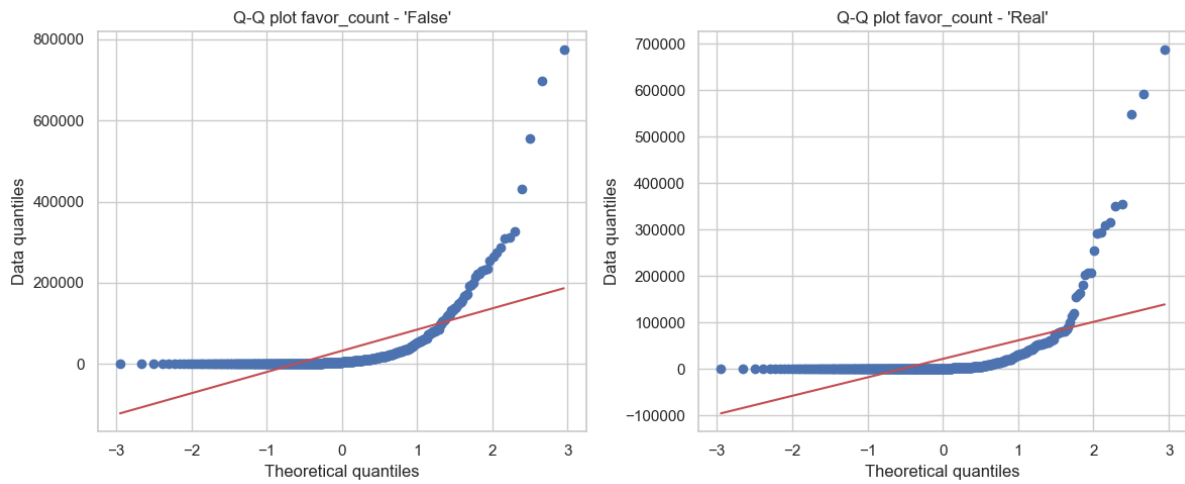
Appendix 7: Normal quantile-quantile plots of users' registration time by label



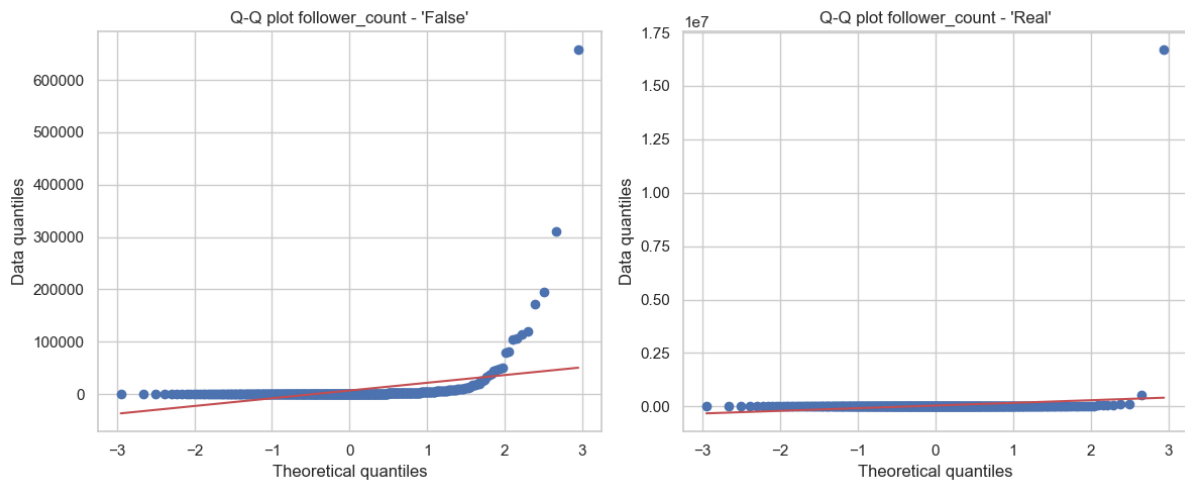
Appendix 8: Normal quantile-quantile plots of users' number of tweets by label



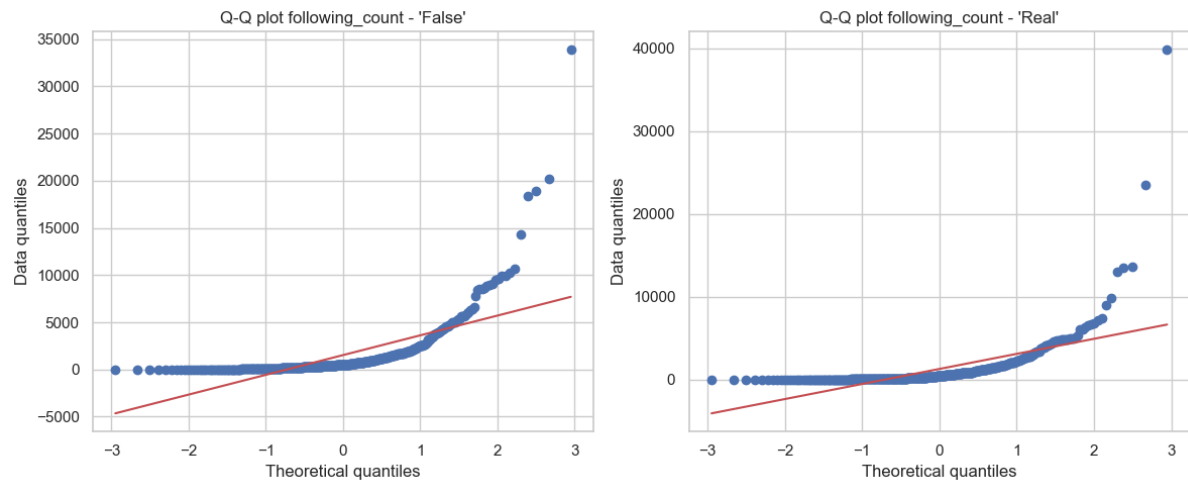
Appendix 9: Normal quantile-quantile plots of the number of tweets liked per user by label



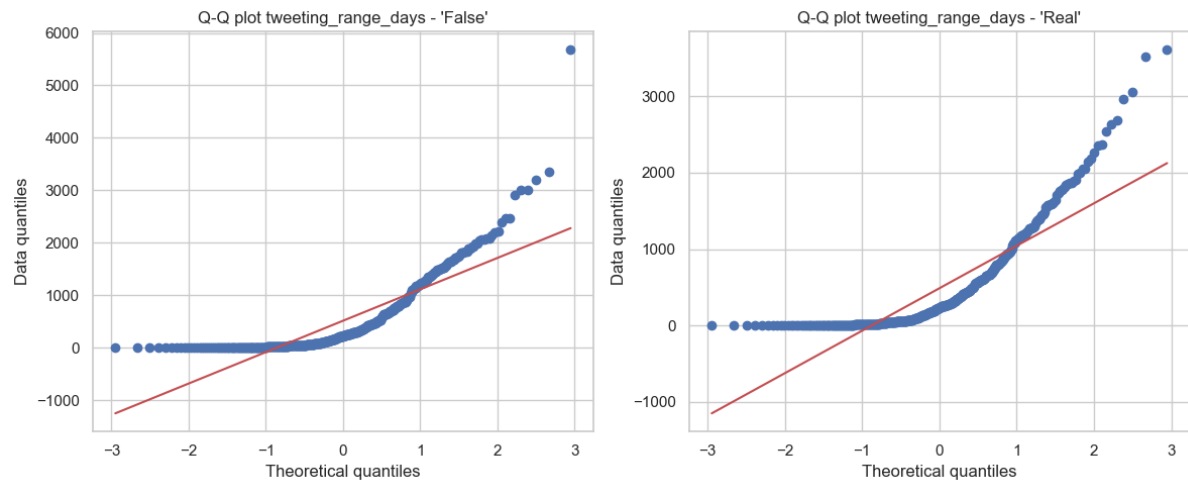
Appendix 10: Normal quantile-quantile plots of users' number of followers by label



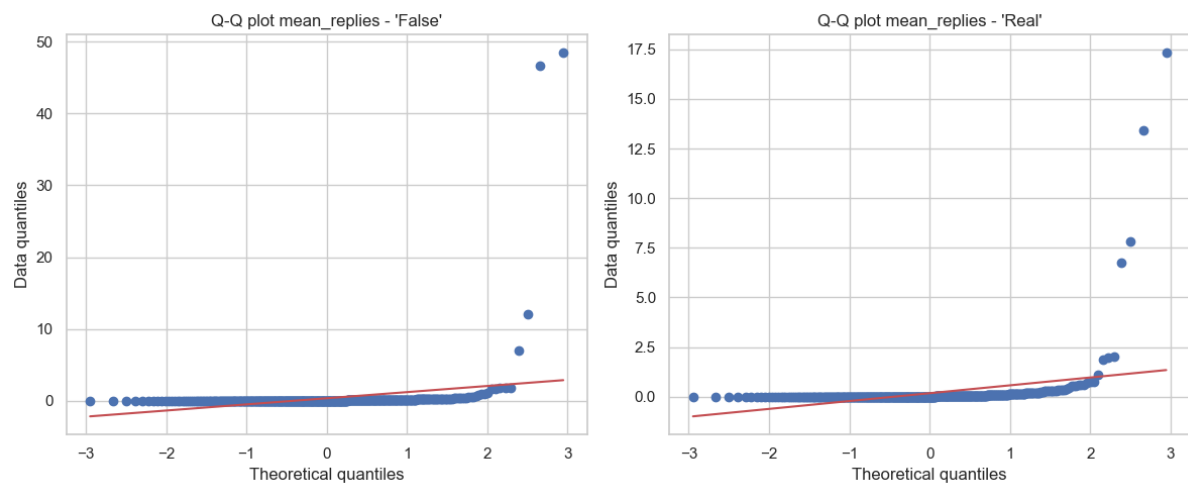
Appendix 11: Normal quantile-quantile plots of users' number of followings by label



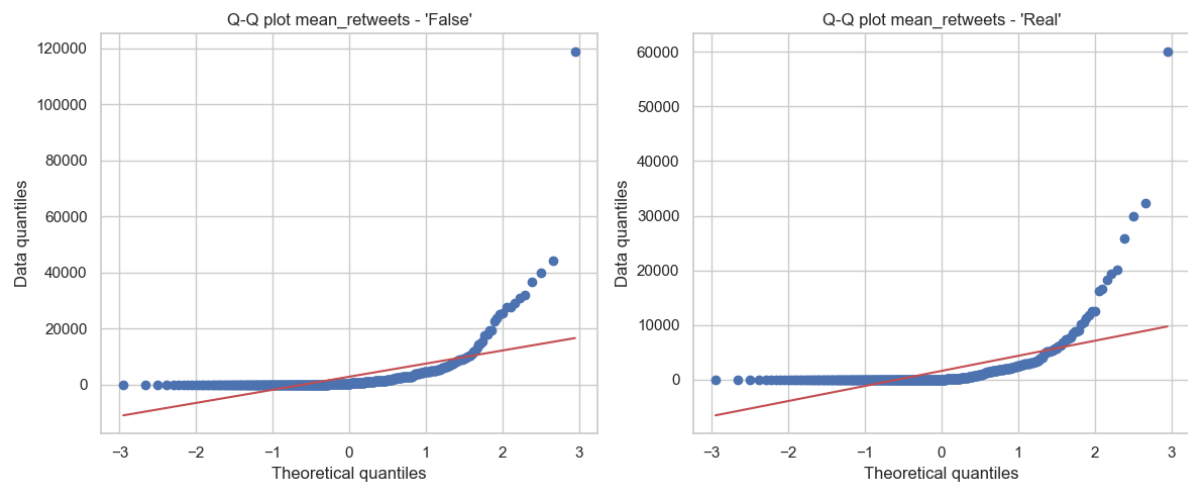
Appendix 12: Normal quantile-quantile plots of users' tweeting time range by label



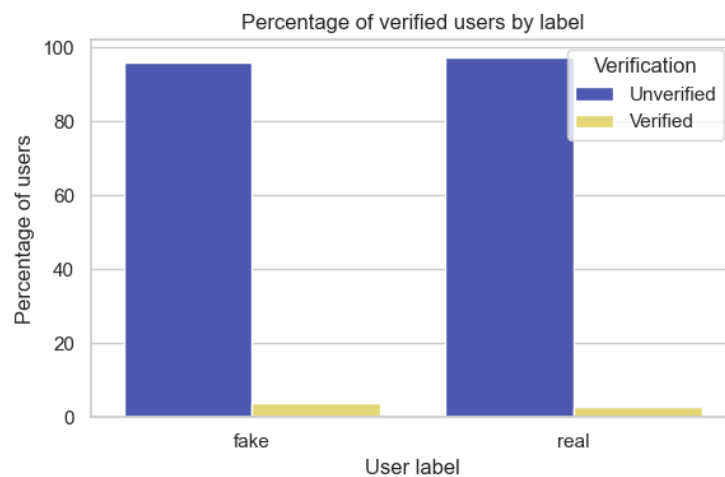
Appendix 13: Normal quantile-quantile plots of the average number of replies generated per user by label



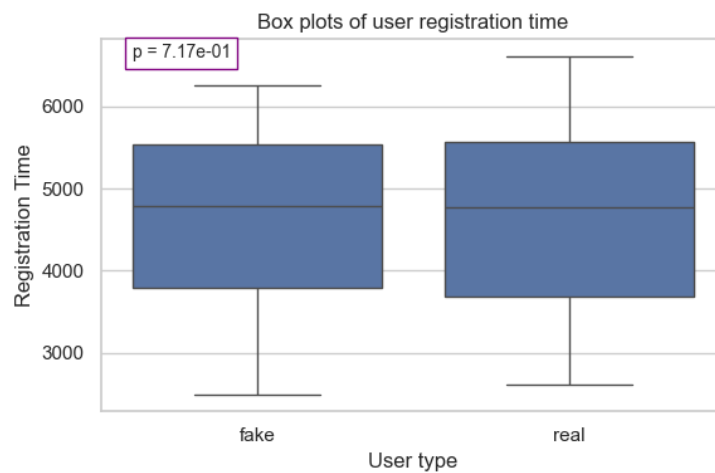
Appendix 14: Normal quantile-quantile plots of the average number of retweets generated per user by label



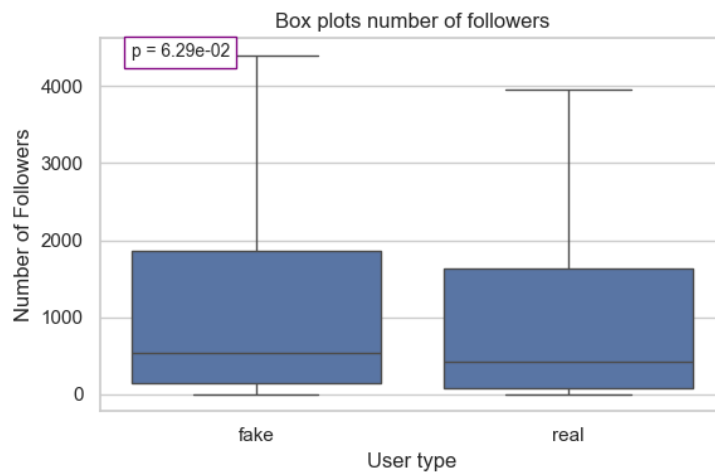
Appendix 15: Percentage of verified users by label



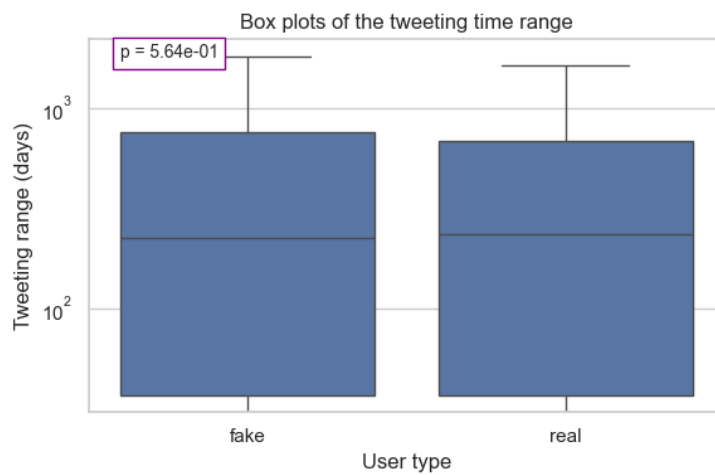
Appendix 16: Box plots of user registration time



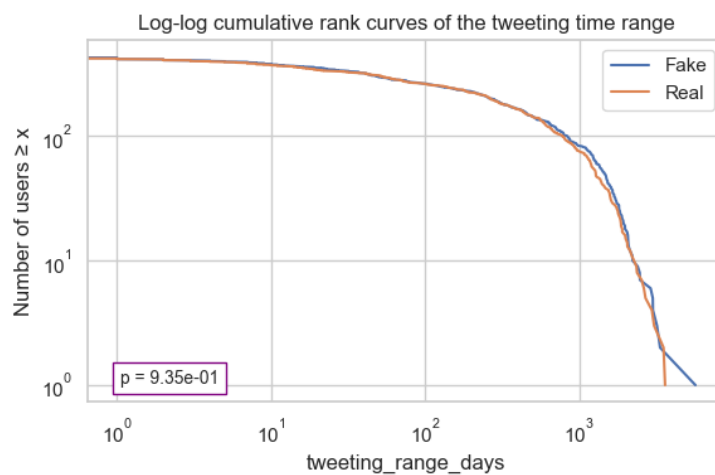
Appendix 17: Box plots of the number of followers



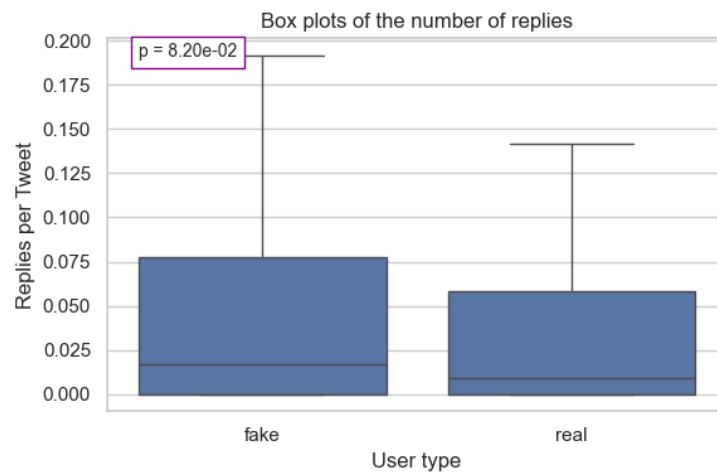
Appendix 18: Box plots of the tweeting time range



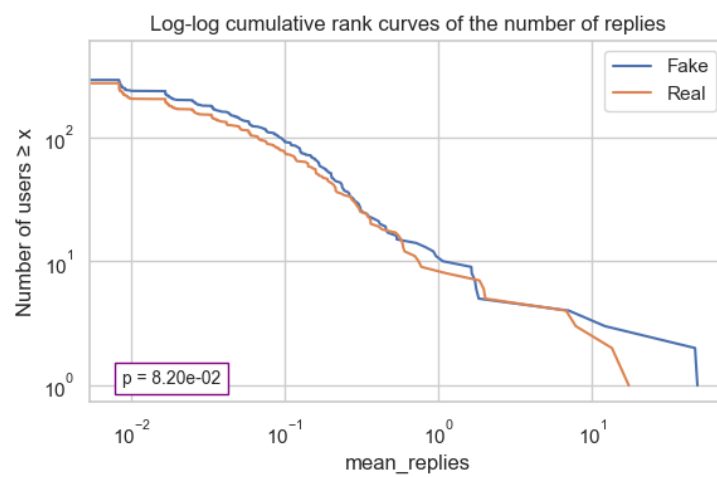
Appendix 19: Log-log cumulative rank curves of the tweeting time range



Appendix 20: Box plots of the number of replies



Appendix 21: Log-log cumulative rank curves of the number of replies



8. References

- Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 734–749. <https://doi.org/10.1109/TKDE.2005.99>
- Aggarwal, C. (2016). Recommender Systems: The Textbook. Springer. <https://link.springer.com/book/10.1007/978-3-319-29659-3>
- Ahlers, D. (2006). News Consumption and the New Electronic Media. *Harvard International Journal of Press/Politics*, 11(1), 29–52. <https://doi.org/10.1177/1081180X05284317>
- Ahmad, I., Yousaf, M., Yousaf, S., Ahmad, M. O., & Uddin, M. I. (2020). Fake News Detection Using Machine Learning Ensemble Methods. *Complexity*, 2020. <https://doi.org/10.1155/2020/8885861>
- Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2), 211–236. <https://doi.org/10.1257/jep.31.2.211>
- Andrea, C. (2022). *Elgar Encyclopedia of Technology and Politics*. Edward Elgar Publishing.
- Arceneaux, K., & Johnson, M. (2013). *Changing Minds or Changing Channels?: Partisan News in an Age of Choice*. University of Chicago Press.
- Areeb, Q. M., Nadeem, M., Sohail, S. S., Imam, R., Doctor, F., Himeur, Y., Hussain, A., & Amira, A. (2023). Filter bubbles in recommender systems: Fact or fallacy—A systematic review. *WIREs Data Mining and Knowledge Discovery*, 13(6), e1512. <https://doi.org/10.1002/widm.1512>
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds), *Proceedings of the Seventh International Conference on Language Resources and Evaluation*

- (LREC'10). European Language Resources Association (ELRA).
<https://aclanthology.org/L10-1531/>
- Bache, K., Newman, D., & Smyth, P. (2013). Text-based measures of document diversity. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 23–31. <https://doi.org/10.1145/2487575.2487672>
- Baden, C., & Springer, N. (2015). Conceptualizing viewpoint diversity in news discourse. *Journalism*, 18(2), 176-194. <https://doi.org/10.1177/1464884915605028>
- Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132. <https://doi.org/10.1126/science.aaa1160>
- Barabási, & Albert-László. (2002). Linked: The New Science of Networks. *American Journal of Physics*, 71(4). <https://doi.org/10.1119/1.1538577>
- Beam, M. A. (2014). Automating the News: How Personalized News Recommender System Design Choices Impact News Reception. *Communication Research*, 41(8), 1019–1041. <https://doi.org/10.1177/0093650213497979>
- Bennett, W. L. (1996). An introduction to journalism norms and representations of politics. *Political Communication*, 13(4), 373–384. <https://doi.org/10.1080/10584609.1996.9963126>
- Berger, J., & Milkman, K. L. (2012). What Makes Online Content Viral? *Journal of Marketing Research*, 49(2), 192–205. <https://doi.org/10.1509/jmr.10.0353>
- Berkovsky, S., Kuflik, T., & Ricci, F. (2008). Mediation of user models for enhanced personalization in recommender systems. *User Modeling and User-Adapted Interaction*, 18(3), 245–286. <https://doi.org/10.1007/s11257-007-9042-9>

- Berkovsky, S., Kuflik, T., & Ricci, F. (2009). Cross-representation mediation of user models. *User Modeling and User-Adapted Interaction*, 19(1), 35–63. <https://doi.org/10.1007/s11257-008-9055-z>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3. <https://robotics.stanford.edu/~ang/papers/jair03-lda.pdf>
- Boltanski, L., & Thévenot, L. (2006). On Justification: Economies of Worth. *Princeton University Press*. <https://press.princeton.edu/books/paperback/9780691125169/on-justification>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (No. arXiv:2108.07258). arXiv. <https://doi.org/10.48550/arXiv.2108.07258>
- Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology*, 15(3), 209–227. <https://doi.org/10.1007/s10676-013-9321-6>
- Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4), 331–370. <https://doi.org/10.1023/A:1021240730564>
- Calvillo, D. P., & Smelter, T. J. (2020). An initial accuracy focus reduces the effect of prior exposure on perceived accuracy of news headlines. *Cognitive Research: Principles and Implications*, 5(1), 55. <https://doi.org/10.1186/s41235-020-00257-y>
- Cambria, E., Zhang, X., Mao, R., Chen, M., & Kwok, K. (2024). SenticNet 8: Fusing Emotion AI and Commonsense AI for Interpretable, Trustworthy, and Explainable Affective Computing. In H. Degen & S. Ntoa (Eds), *HCI International 2024 – Late Breaking*

- Papers* (Vol. 15382, pp. 197–216). Springer Nature Switzerland.
https://doi.org/10.1007/978-3-031-76827-9_11
- Carlson, M. (2007). Order versus access: News search engines and the challenge to traditional journalistic roles. *Media, Culture & Society*, 29(6), 1014–1030.
<https://doi.org/10.1177/0163443707084346>
- Chen, K., Yang, S., & Duan, Z. (2021). Twitter as Research Data: Tools, Costs, Skillsets, and Lessons Learned. *Politics and the Life Sciences*, 41(1), 1–23.
<https://doi.org/10.1017/pls.2021.19>
- Chueca Del Cerro, C. (2024). The power of social networks and social media’s filter bubble in shaping polarisation: An agent-based model. *Applied Network Science*, 9(1), 69.
<https://doi.org/10.1007/s41109-024-00679-3>
- 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>
- Cinelli, M., Morales, G. D. F., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2020). *Echo Chambers on Social Media: A comparative analysis* (No. arXiv:2004.09603). arXiv.
<https://doi.org/10.48550/arXiv.2004.09603>
- Cordonier, L., & Brest, A. (2021). *How do the French inform themselves on the Internet? Analysis of online information and disinformation behaviors* [Research Report]. Fondation Descartes. <https://hal.science/hal-03167734>
- Cotton, J. L. (1985). Cognitive Dissonance in Selective Exposure. In *Selective Exposure To Communication*. Routledge.
- Dahlgren, P. M. (2021). A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review*, 42(1), 15–33. <https://doi.org/10.2478/nor-2021-0002>

- D'Alessio, D., & Allen, M. (2007). The Selective Exposure Hypothesis and Media Choice Processes. In *Mass media effects research: Advances through meta-analysis* (pp. 103–118). Lawrence Erlbaum Associates Publishers.
- Davies, H. C. (2018). Redefining Filter Bubbles as (Escapable) Socio-Technical Recursion. *Sociological Research Online*, 23(3), 637–654. <https://doi.org/10.1177/1360780418763824>
- Del Vicario, M., Scala, A., Caldarelli, G., Stanley, H. E., & Quattrociocchi, W. (2017). Modeling confirmation bias and polarization. *Scientific Reports*, 7(1), 40391. <https://doi.org/10.1038/srep40391>
- Diaz Ruiz, C., & Nilsson, T. (2023). Disinformation and Echo Chambers: How Disinformation Circulates on Social Media Through Identity-Driven Controversies. *Journal of Public Policy & Marketing*, 42(1), 18–35. <https://doi.org/10.1177/07439156221103852>
- Donati, P. (1992). Political discourse analysis. In: Diani M and Eyerman R (eds). *Studying Collective Action*, 671–686.
- Du, K. (2022, March 7). *Evaluating Hyperparameter Alpha of LDA Topic Modeling*. DHd 2022 Kulturen des digitalen Gedächtnisses. 8. Tagung des Verbands ‘Digital Humanities im deutschsprachigen Raum’ (DHd 2022), Potsdam. <https://doi.org/10.5281/zenodo.6327965>
- Eady, G., Nagler, J., Guess, A., Zilinsky, J., & Tucker, J. A. (2019). How Many People Live in Political Bubbles on Social Media? Evidence From Linked Survey and Twitter Data. *SAGE Open*, 9(1), 2158244019832705. <https://doi.org/10.1177/2158244019832705>
- Entman, R. (1993). Framing: Toward Clarification of A Fractured Paradigm. *The Journal of Communication*, 43, 51–58. <https://doi.org/10.1111/j.1460-2466.1993.tb01304.x>

- Entman, R. M., & Wildman, S. S. (1992). Reconciling economic and non-economic perspectives on media policy: Transcending the “marketplace of ideas.” *Journal of Communication*, 42(1), 5–19. <https://doi.org/10.1111/j.1460-2466.1992.tb00765.x>
- Fairclough, N. (2003). *Analysing Discourse: Textual Analysis for Social Research* (1st edn). Routledge. <https://doi.org/10.4324/9780203697078>
- Fischer, P., Lea, S., Kastenmüller, A., Greitemeyer, T., Fischer, J., & Frey, D. (2011). The process of selective exposure: Why confirmatory information search weakens over time. *Organizational Behavior and Human Decision Processes*, 114(1), 37–48. <https://doi.org/10.1016/j.obhdp.2010.09.001>
- Flanagin, A. J., & Metzger, M. J. (2000). Perceptions of Internet Information Credibility. *Journalism & Mass Communication Quarterly*, 77(3), 515–540. <https://doi.org/10.1177/107769900007700304>
- Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, 80(S1), 298–320. <https://doi.org/10.1093/poq/nfw006>
- Fleder, D., & Hosanagar, K. (2009). Blockbuster Culture’s Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science*, 55(5), 697–712. <https://doi.org/10.1287/mnsc.1080.0974>
- Fletcher, R., & Nielsen, R. K. (2017). Are News Audiences Increasingly Fragmented? A Cross-National Comparative Analysis of Cross-Platform News Audience Fragmentation and Duplication. *Journal of Communication*, 67(4), 476–498. <https://doi.org/10.1111/jcom.12315>
- Fletcher, R., & Nielsen, R. K. (2018). Are people incidentally exposed to news on social media? A comparative analysis. *New Media & Society*, 20(7), 2450–2468. <https://doi.org/10.1177/1461444817724170>

- Frey, D. (1986). Recent Research on Selective Exposure to Information. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 19, pp. 41–80). Academic Press.
[https://doi.org/10.1016/S0065-2601\(08\)60212-9](https://doi.org/10.1016/S0065-2601(08)60212-9)
- Garrett, R. K. (2009). Politically Motivated Reinforcement Seeking: Reframing the Selective Exposure Debate. *Journal of Communication*, 59(4), 676–699.
<https://doi.org/10.1111/j.1460-2466.2009.01452.x>
- Geetha, G., Safa, M., Fancy, C., & Saranya, D. (2018). A Hybrid Approach using Collaborative filtering and Content based Filtering for Recommender System. *Journal of Physics: Conference Series*, 1000(1), 012101. <https://doi.org/10.1088/1742-6596/1000/1/012101>
- Gelfert, A. (2018). Fake News: A Definition. *Informal Logic*, 38(1), 84–117.
<https://doi.org/10.22329/il.v38i1.5068>
- George, C. P. (2015). *Latent Dirichlet Allocation: Hyperparameter Selection and Applications to Electronic Discovery*. University of Florida.
- Gharahighehi, A., & Vens, C. (2023). Diversification in session-based news recommender systems. *Personal and Ubiquitous Computing*, 27(1), 5–15.
<https://doi.org/10.1007/s00779-021-01606-4>
- Gomez-Uribe, C. A., & Hunt, N. (2016). The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manage. Inf. Syst.*, 6(4), 13:1-13:19.
<https://doi.org/10.1145/2843948>
- Graham, G. (2017). Behaviorism. In E. N. Zalta & U. Nodelman (Eds), *The Stanford Encyclopedia of Philosophy* (Spring 2017). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2017/entries/behaviorism/>

- Greifeneder, R., Jaffe, M., Newman, E., & Schwarz, N. (Eds). (2021). *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation*. Routledge.
<https://doi.org/10.4324/9780429295379>
- Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, 4(5), 472–480. <https://doi.org/10.1038/s41562-020-0833-x>
- Haim, M., Graefe, A., & Brosius, H.-B. (2017). Burst of the Filter Bubble?: Effects of personalization on the diversity of Google News. *Digital Journalism*.
<https://doi.org/10.1080/21670811.2017.1338145>
- Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. *Elsevier*, 3rd edition.
- Hansson, S. O., & Grüne-Yanoff, T. (2012). Preferences. In E. N. Zalta & U. Nodelman (Eds), *The Stanford Encyclopedia of Philosophy* (Winter 2012). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2012/entries/preferences/>
- Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4), 555–588.
<https://doi.org/10.1037/a0015701>
- Hart, W., Richardson, K., Tortoriello, G. K., & Earl, A. (2020). ‘You Are What You Read:’ Is selective exposure a way people tell us who they are? *British Journal of Psychology*, 111(3), 417–442. <https://doi.org/10.1111/bjop.12414>
- Helberger, N., Karppinen, Kari, & D’Acunzio, L. (2018). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2), 191–207. <https://doi.org/10.1080/1369118X.2016.1271900>

- Holan, A. D. (2016). *PolitiFact's 2016 Lie of the Year: Fake news*. Politifact.
<https://www.politifact.com/article/2016/dec/13/2016-lie-year-fake-news/>
- HosseiniMotlagh, S., & Papalexakis, E. E. (2018, February). Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles. *Misinformation and Misbehavior Mining on the Web Workshop*.
<https://www.semanticscholar.org/paper/Unsupervised-Content-Based-Identification-of-Fake-Papalexakis/98e95b2af4de42f9155b285c4527e5492e367c83>
- 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>
- Kanoh, H. (2018). Why do people believe in fake news over the Internet? An understanding from the perspective of existence of the habit of eating and drinking. *Procedia Computer Science*, 126, 1704–1709. <https://doi.org/10.1016/j.procs.2018.08.107>
- Karlsen, R., Steen-Johnsen, K., Wollebæk, D., & Enjolras, B. (2017). Echo chamber and trench warfare dynamics in online debates. *European Journal of Communication*, 32(3), 257–273. <https://doi.org/10.1177/0267323117695734>
- Kim, B., Xiong, A., Lee, D., & Han, K. (2021). A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *PLoS ONE*, 16(12), e0260080.
<https://doi.org/10.1371/journal.pone.0260080>
- Kim, H.-N., Ha, I., Lee, K.-S., Jo, G.-S., & El-Saddik, A. (2011). Collaborative user modeling for enhanced content filtering in recommender systems. *Decision Support Systems*, 51(4), 772–781. <https://doi.org/10.1016/j.dss.2011.01.012>
- Kogan, S., Moskowitz, T. J., & Niessner, M. (2018). Fake News: Evidence from Financial Markets. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3237763>

- Koren, Y., Rendle, S., & Bell, R. (2022). Advances in Collaborative Filtering. In F. Ricci, L. Rokach, & B. Shapira (Eds), *Recommender Systems Handbook* (pp. 91–142). Springer US. https://doi.org/10.1007/978-1-0716-2197-4_3
- Kucharski, A. (2016). Study epidemiology of fake news. *Nature*, 540(7634), 525–525. <https://doi.org/10.1038/540525a>
- Kumar, B., & Sharma, N. (2016). Approaches, Issues and Challenges in Recommender Systems: A Systematic Review. *Indian Journal of Science and Technology*, 9(1), 1–12. <https://doi.org/10.17485/ijst/2016/v9i47/94892>
- Lahoti, P., Garimella, K., & Gionis, A. (2017). *Joint Non-negative Matrix Factorization for Learning Ideological Leaning on Twitter* (No. arXiv:1711.10251). arXiv. <https://doi.org/10.48550/arXiv.1711.10251>
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380), 1094–1096. <https://doi.org/10.1126/science.aao2998>
- Lee, D., & Hosanagar, K. (2014). *Impact of Recommender Systems on Sales Volume and Diversity*.
- Leydesdorff, L. (2018). Diversity and interdisciplinarity: How can one distinguish and recombine disparity, variety, and balance? *Scientometrics*, 116(3), 2113–2121. <https://doi.org/10.1007/s11192-018-2810-y>
- Liu, J., Jiang, Y., Li, Z., Zhang, X., & Lu, H. (2016). Domain-Sensitive Recommendation with User-Item Subgroup Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(4), 939–950. <https://doi.org/10.1109/TKDE.2015.2492540>
- Lü, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K., & Zhou, T. (2012). Recommender systems. *Physics Reports*, 519(1), 1–49. <https://doi.org/10.1016/j.physrep.2012.02.006>

- Machete, P., & Turpin, M. (2020). The Use of Critical Thinking to Identify Fake News: A Systematic Literature Review. In M. Hattingh, M. Matthee, H. Smuts, I. Pappas, Y. K. Dwivedi, & M. Mäntymäki (Eds), *Responsible Design, Implementation and Use of Information and Communication Technology* (pp. 235–246). Springer International Publishing. https://doi.org/10.1007/978-3-030-45002-1_20
- Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, 5(1), 47. <https://doi.org/10.1186/s41235-020-00252-3>
- McQuail, D. (1993). Media Performance: Mass Communication and the Public Interest. *Canadian Journal of Communication*, 18(4). <https://doi.org/10.22230/cjc.1993v18n4a783>
- Meadows, D. H. (2008). *Thinking in Systems: A Primer*. Chelsea Green Publishing.
- Meel, P., & Vishwakarma, D. K. (2019). Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities. *Expert Systems with Applications*, 153, 112986. <https://doi.org/10.1016/j.eswa.2019.112986>
- Melki, J., Tamim, H., Hadid, D., Makki, M., Amine, J. E., & Hitti, E. (2021). Mitigating infodemics: The relationship between news exposure and trust and belief in COVID-19 fake news and social media spreading. *PLoS ONE*, 16(6), e0252830. <https://doi.org/10.1371/journal.pone.0252830>
- Melville, P., Mooney, R. J., & Nagarajan, R. (2002). Content-Boosted Collaborative Filtering for Improved Recommendations. *Proceedings of the 18th National Conference on Artificial Intelligence*, 187–192.
- Melville, P., & Sindhvani, V. (2011). *Recommender Systems* (Springer US). <https://www.prem-melville.com/publications/recommender-systems-eml2010.pdf>

- Messing, S., & Westwood, S. J. (2014). Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. *Communication Research*, 41(8), 1042–1063. <https://doi.org/10.1177/0093650212466406>
- Michael, R. B., & Breaux, B. O. (2021). The relationship between political affiliation and beliefs about sources of “fake news”. *Cognitive Research: Principles and Implications*, 6(1), 6. <https://doi.org/10.1186/s41235-021-00278-1>
- Mohanty, S., Chatterjee, J., Jain, S., Elngar, A., & Gupta, P. (2020). *Recommender System with Machine Learning and Artificial Intelligence*. <https://doi.org/10.1002/9781119711582>
- Mohd Razali, N., & Yap, B. (2011). Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *J. Stat. Model. Analytics*, 2(1), 21–33.
- Muñoz, P., Díez, F., & Bellogín, A. (2024). Modeling disinformation networks on Twitter: Structure, behavior, and impact. *Applied Network Science*, 9(1), 4. <https://doi.org/10.1007/s41109-024-00610-w>
- Munson, S. A., & Resnick, P. (2010). Presenting diverse political opinions: How and how much. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1457–1466. <https://doi.org/10.1145/1753326.1753543>
- Mutz, R. (2022). Diversity and interdisciplinarity: Should variety, balance and disparity be combined as a product or better as a sum? An information-theoretical and statistical estimation approach. *Scientometrics*, 127(12), 7397–7414. <https://doi.org/10.1007/s11192-022-04336-3>
- Nadarevic, L., Reber, R., Helmecke, A. J., & Köse, D. (2020). Perceived truth of statements and simulated social media postings: An experimental investigation of source credibility, repeated exposure, and presentation format. *Cognitive Research: Principles and Implications*, 5(1), 56. <https://doi.org/10.1186/s41235-020-00251-4>

- Narayanan, A. (2023). *Understanding Social Media Recommendation Algorithms*.
<https://doi.org/10.7916/khdk-m460>
- Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D. A. L., & Nielsen, R. K. (2017). Reuters Institute Digital News Report 2017. *Oxford: Reuters Institute for the Study of Journalism*.
- 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>
- Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2, 175.
- Oechslein, O., & Hess, T. (2013). Incorporating Social Networking Information in Recommender Systems: The Development of a Classification Framework. *BLED 2013 Proceedings*. <https://aisel.aisnet.org/bled2013/19>
- Pandarachalil, R., Sendhilkumar, S., & Mahalakshmi, G. S. (2015). Twitter Sentiment Analysis for Large-Scale Data: An Unsupervised Approach. *Cognitive Computation*, 7(2), 254–262. <https://doi.org/10.1007/s12559-014-9310-z>
- Pariser, E. (2011). *The Filter Bubble: What The Internet Is Hiding From You*. Penguin UK.
- Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press.
<https://doi.org/10.4159/harvard.9780674736061>
- Pehlivanoglu, D., Lin, T., Deceus, F., Heemskerk, A., Ebner, N. C., & Cahill, B. S. (2021). The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. *Cognitive Research: Principles and Implications*, 6(1), 24.
<https://doi.org/10.1186/s41235-021-00292-3>

- Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12), 1865–1880. <https://doi.org/10.1037/xge0000465>
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*, 31(7), 770–780.
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>
- Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, 88(2), 185–200. <https://doi.org/10.1111/jopy.12476>
- Pennycook, G., & Rand, D. G. (2021). The Psychology of Fake News. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
- Pfeffer, J., Matter, D., Jaidka, K., Varol, O., Mashhadi, A., Lasser, J., Assenmacher, D., Wu, S., Yang, D., Brantner, C., Romero, D. M., Otterbacher, J., Schwemmer, C., Joseph, K., Garcia, D., & Morstatter, F. (2023). *Just Another Day on Twitter: A Complete 24 Hours of Twitter Data* (No. arXiv:2301.11429). arXiv. <https://doi.org/10.48550/arXiv.2301.11429>
- Pierri, F., Piccardi, C., & Ceri, S. (2020). Topology comparison of Twitter diffusion networks effectively reveals misleading information. *Scientific Reports*, 10(1), 1372. <https://doi.org/10.1038/s41598-020-58166-5>
- Preston, S., Anderson, A., Robertson, D. J., Shephard, M. P., & Huhe, N. (2021). Detecting fake news on Facebook: The role of emotional intelligence. *PLoS ONE*, 16(3), e0246757. <https://doi.org/10.1371/journal.pone.0246757>

- Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1), 24–43. [https://doi.org/10.1016/0040-5809\(82\)90004-1](https://doi.org/10.1016/0040-5809(82)90004-1)
- Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: An open architecture for collaborative filtering of netnews. *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, 175–186. <https://doi.org/10.1145/192844.192905>
- Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to Recommender Systems Handbook. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds), *Recommender Systems Handbook* (pp. 1–35). Springer US. https://doi.org/10.1007/978-0-387-85820-3_1
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. WSDM 2015: Eighth ACM International Conference on Web Search and Data Mining. <https://doi.org/10.1145/2684822.2685324>
- Ross, L., & Ward, A. (1996). Naive Realism in Everyday Life: Implications for Social Conflict and Misunderstanding. In *Values and Knowledge*. Psychology Press.
- Rowland, F. (2011). The Filter Bubble: What the Internet is Hiding from You (review). *Portal: Libraries and the Academy*, 11(4), 1009–1011.
- Salovey, P., & Mayer, J. D. (1990). Emotional Intelligence. *Imagination, Cognition and Personality*, 9(3), 185–211. <https://doi.org/10.2190/DUGG-P24E-52WK-6CDG>
- Salter, J., & Antonopoulos, N. (2006). CinemaScreen recommender agent: Combining collaborative and content-based filtering. *IEEE Journals & Magazine*, 21, 35–41.
- Samuels, M. G. (2012). Review: The Filter Bubble: What the Internet is Hiding from You by Eli Pariser. *InterActions: UCLA Journal of Education and Information Studies*, 8(2). <https://doi.org/10.5070/d482011835>

- Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. *The Adaptive Web: Methods and Strategies of Web Personalization*. https://link.springer.com/content/pdf/10.1007/978-3-540-72079-9_9?pdf=chapter%20toc
- Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2001). Generative Models for Cold-Start Recommendations. *Proceedings of the 2001 SIGIR Workshop on Recommender Systems*, 6.
- Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 253–260. <https://doi.org/10.1145/564376.564421>
- Schomakers, E.-M., Lidynia, C., & Ziefle, M. (2020). All of me? Users' preferences for privacy-preserving data markets and the importance of anonymity. *Electronic Markets*, 30(3), 649–665. <https://doi.org/10.1007/s12525-020-00404-9>
- Shin, Y., Sojdehei, Y., Zheng, L., & Blanchard, B. (2023). Content-Based Unsupervised Fake News Detection on Ukraine-Russia War. *SMU Data Science Review*, 7(1). <https://scholar.smu.edu/datasciencereview/vol7/iss1/3>
- Shu, K., Bernard, H. R., & Liu, H. (2019). Studying Fake News via Network Analysis: Detection and Mitigation. In N. Agarwal, N. Dokoohaki, & S. Tokdemir (Eds), *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining* (pp. 43–65). Springer International Publishing. https://doi.org/10.1007/978-3-319-94105-9_3
- Shu, K., Mahudeswaran, D., Wang, S., Lee, D., & Liu, H. (2019). *FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for*

- Studying Fake News on Social Media* (No. arXiv:1809.01286). arXiv.
<https://doi.org/10.48550/arXiv.1809.01286>
- Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake News Detection on Social Media: A Data Mining Perspective. *SIGKDD Explor. Newsl.*, 19(1), 22–36.
<https://doi.org/10.1145/3137597.3137600>
- Shu, K., Wang, S., & Liu, H. (2017). *Exploiting Tri-Relationship for Fake News Detection*.
<https://doi.org/10.48550/arXiv.1712.07709>
- Shu, K., Wang, S., & Liu, H. (2018). Understanding User Profiles on Social Media for Fake News Detection. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 430–435. <https://doi.org/10.1109/MIPR.2018.00092>
- Sindermann, C., Schmitt, H. S., Rozgonjuk, D., Elhai, J. D., & Montag, C. (2021). The evaluation of fake and true news: On the role of intelligence, personality, interpersonal trust, ideological attitudes, and news consumption. *Heliyon*, 7(3).
<https://doi.org/10.1016/j.heliyon.2021.e06503>
- Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), 127:1-127:26. <https://doi.org/10.1145/3359229>
- Stein, J., Keuschnigg, M., & van de Rijt, A. (2023). Network segregation and the propagation of misinformation. *Scientific Reports*, 13(1), 917. <https://doi.org/10.1038/s41598-022-26913-5>
- Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15), 707–719.
<https://doi.org/10.1098/rsif.2007.0213>

- Strömbäck, J. (2005). In Search of a Standard: Four models of democracy and their normative implications for journalism. *Journalism Studies*, 6(3), 331–345.
<https://doi.org/10.1080/14616700500131950>
- Stroud, N. J. (2008). Media Use and Political Predispositions: Revisiting the Concept of Selective Exposure. *Political Behavior*, 30(3), 341–366.
<https://doi.org/10.1007/s11109-007-9050-9>
- Su, X., & Khoshgoftaar, T. M. (2009). A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, 2009(1), 421425.
<https://doi.org/10.1155/2009/421425>
- Sülflow, M., Schäfer, S., & Winter, S. (2019). Selective attention in the news feed: An eye-tracking study on the perception and selection of political news posts on Facebook. *New Media & Society*, 21(1), 168–190. <https://doi.org/10.1177/1461444818791520>
- Sundar, S. S., & Marathe, S. S. (2010). Personalization versus Customization: The Importance of Agency, Privacy, and Power Usage. *Human Communication Research*, 36(3), 298–322. <https://doi.org/10.1111/j.1468-2958.2010.01377.x>
- Sunstein, C. R. (2001). *Republic.com* (Princeton University Press).
- Tat, S., Myra, E., Ferree, M., Gamson, W., Gerhards, J., & Rucht, D. (2002). *Shaping Abortion Discourse*. <https://doi.org/10.1017/CBO9780511613685>
- Tetlock, P. E. (2017). *Expert Political Judgment: How Good Is It? How Can We Know? - New Edition*. Princeton University Press.
<https://www.torrossa.com/en/resources/an/5560194>
- Tharenou, P. (with Cooper, B., & Donohue, R.). (2007). *Management research methods*. Cambridge University Press. <http://hdl.handle.net/2078/ebook:330917>

- Thurman, N. (2011). *Making 'The Daily Me': Technology, economics and habit in the mainstream assimilation of personalized news*.
<https://journals.sagepub.com/doi/10.1177/1464884910388228>
- Thurman, N., & Schifferes, S. (2012). The Future of Personalization at News Websites: Lessons from a longitudinal study. *Journalism Studies*, 13(5–6), 775–790.
<https://doi.org/10.1080/1461670X.2012.664341>
- Tsfati, Y., Boomgaarden, H. G., Strömbäck, J., Vliegenthart, R., Damstra, A., & Lindgren, E. (2020). Causes and Consequences of Mainstream Media Dissemination of Fake News: Literature Review and Synthesis. *Annals of the International Communication Association*, 44(2), 157–173. <https://doi.org/10.1080/23808985.2020.1759443>
- van Dam, A. (2019). Diversity and its decomposition into variety, balance and disparity. *Royal Society Open Science*, 6(7), 190452. <https://doi.org/10.1098/rsos.190452>
- Voakes, P. S., Kapfer, J., Kurpius, D., & Chern, D. S.-Y. (1996). Diversity in the News: A Conceptual and Methodological Framework. *Journalism & Mass Communication Quarterly*, 73(3), 582–593. <https://doi.org/10.1177/107769909607300306>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Vraga, E., Bode, L., & Troller-Renfree, S. (2016). Beyond Self-Reports: Using Eye Tracking to Measure Topic and Style Differences in Attention to Social Media Content. *Communication Methods and Measures*, 10(2–3), 149–164.
<https://doi.org/10.1080/19312458.2016.1150443>
- Wang, C.-C. (2020). Fake News and Related Concepts: Definitions and Recent Research Development. *Contemporary Management Research*, 16(3), Article 3.
<https://doi.org/10.7903/cmr.20677>

- Wang, Y., Guo, J., Yuan, C., & Li, B. (2022). Sentiment Analysis of Twitter Data. *Applied Sciences*, 12(22), Article 22. <https://doi.org/10.3390/app122211775>
- Webster, J. G. (2017). Three myths of digital media. *Convergence*, 23(4), 352–361. <https://doi.org/10.1177/1354856517700385>
- Weeks, B. E., Ksiazek, T. B., & Holbert, R. L. (2016). Partisan Enclaves or Shared Media Experiences? A Network Approach to Understanding Citizens' Political News Environments. *Journal of Broadcasting & Electronic Media*, 60(2), 248–268. <https://doi.org/10.1080/08838151.2016.1164170>
- Wetherell, M., & Potter, J. (1988). Discourse analysis and the identification of interpretative repertoires. *Analysing Everyday Explanation: A Casebook of Methods*, 1688183.
- Winters, R., Winters, A., & Amedee, R. G. (2010). Statistics: A Brief Overview. *Ochsner Journal*, 10(3), 213–216.
- Wogalter, M. S., & Mayhorn, C. B. (2008). Trusting the Internet: Cues Affecting Perceived Credibility. *International Journal of Technology and Human Interaction (IJTHI)*, 4(1), 75–93. <https://doi.org/10.4018/jthi.2008010105>
- Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., & Liu, H. (2019). Unsupervised fake news detection on social media: A generative approach. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 33, 5644–5651. <https://doi.org/10.1609/aaai.v33i01.33015644>
- Zeng, J., Chan, C. H., & Fu, K. W. (2021). How social media echo chambers and algorithms influence misinformation diffusion. *Computers in Human Behavior Reports*, 4, 100126.
- Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2), 102025. <https://doi.org/10.1016/j.ipm.2019.03.004>

- Zhou, X., & Zafarani, R. (2020). A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys*, 53(5), 109:1-109:40.
<https://doi.org/10.1145/3395046>
- Zimmer, F., Scheibe, K., & Stock, W. (2019, January 3). Echo Chambers and Filter Bubbles of Fake News in Social Media. Man-made or produced by algorithms? *8th Annual Arts, Humanities, Social Sciences & Education Conference*.
- Zuiderveen Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., de Vreese, C., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, 5(1).
<https://doi.org/10.14763/2016.1.401>