DG HOME

EU Internet Forum: Study on the Role and Effects of the Use of Algorithmic Amplification to Spread Terrorist, Violent Extremist and Borderline Content

Final Report

Study commissioned by DG HOME,
drafted by Fincons, Trust Lab and Tremau,
in the context of the EU Internet Forum.
October – 2023

EUROPEAN COMMISSION

Directorate-General for Migration and Home Affairs
Internal Security Unit

Prevention of Radicalisation (HOME.D.3)

E-mail contact:: HOME-INTERNET-FORUM@ec.europa.eu

European Commission
B-1000 Brussels

DG HOME

EU Internet Forum: Study on the Role and Effects of the Use of Algorithmic Amplification to Spread Terrorist, Violent Extremist and Borderline Content

Final Report

Manuscript completed in October 2023

First edition

This document should not be considered as representative of the European Commission's official position.

Luxembourg: Publications Office of the European Union, 2023

Print	ISBN 978-92-68-08172-3	doi: 10.2837/046873	DR-05-23-377-EN-C
PDF	ISBN 978-92-68-08173-0	doi: 10.2837/259157	DR-05-23-377-EN-N

DG HOME

EU INTERNET FORUM: STUDY ON THE ROLE AND EFFECTS OF THE USE OF ALGORITHMIC AMPLIFICATION TO SPREAD TERRORIST, VIOLENT EXTREMIST AND BORDERLINE CONTENT

European Commission

This is the Final Report of the study commissioned by DG HOME, drafted by Fincons, Trust Lab and Tremau, in the context of the EU Internet Forum.

Abstract

The European Commission (DG HOME) commissioned this study for the European Union Internet Forum. The study examines the degree to which the five leading social media platforms in the EU, and in particular their recommender systems which proactively present content to users, algorithmically amplify terrorist and violent extremist content to online users. The study also examines the extent to which recommender systems amplify Borderline content, such as some forms of hate speech leading to violent extremism, disinformation and other forms of legal, but harmful content, that could lead to the creation of online filter bubbles, and recruitment and radicalisation of online users. The study analyses Facebook, Instagram, TikTok, Twitter and YouTube, covering eight markets/languages in the EU. The study was conducted from July 2022 to May 2023. Prior to this Final Report, the team submitted an Inception Report and two Interim Reports that complement this report.

Executive Summary

The European Commission (DG HOME) commissioned this study, performed by Trust Lab in collaboration with Tremau and coordinated by Fincons Group, which examines the degree to which five leading social media platforms and their recommender systems[1] amplify harmful Terrorist and Violent Extremist (TVE) content and "Borderline" content[2] to online users.

The Report reflects the study of Facebook, Instagram, TikTok, Twitter and YouTube, that operate in eight markets/languages, namely, Arabic, English, French, German, Italian, Polish, Russian and Spanish. The study looked at: (1) the extent to which a platform's recommender system algorithmically amplifies TVE and Borderline content to users, and the impact of the dissemination of such content on a user's journey to radicalisation; and (2) the extent to which a platform's recommender system leads to the creation of online "filter bubbles"[3] and the impact on a user's adoption of violent extremist beliefs.

This study was a collaboration between Trust Lab, Professor Theodoros Evgeniou, Professor of Decision Sciences and Technology Management at INSEAD, who researched the risk of radicalisation linked to the dissemination of TVE content, and Fincons Group which handled project coordination.

[1] Recommender systems, also known as content curation systems, are the systems that prioritise content or make personalised content recommendations to online users. A key component of the recommender system is its recommender algorithm that determines the content a user will be served.

[2] As noted by the EU Internet Forum in its Year in Review 2022: "Borderline content, also known as harmful, but legal content, is content that comes close to infringing on the community guidelines of social media platforms or laws regulating online illegal content. Some of the most common types of Borderline content identified in the EU are anti-establishment/anti-institutions, antisemitic, anti-trans, misogynistic, anti-migrants, racist, or against COVID-19 measures."

[3] "Filter bubble "refers to a homogeneous feed/recommended content, caused by the algorithmic amplification of certain content types on a given platform. The higher proportion of similar content, the deeper/larger the filter bubble.

The European Commission's DG CONNECT has also been involved in the development of the study. The study was conducted from July 2022 to May 2023.

The study methodology is novel and discussed in Section 3.2 of the Report. Internal proprietary platform data (including algorithms), data, bots and Application Programming Interface (API) queries, which are common for this type of study, were not used. Trust Lab worked with human analysts to simulate motivated users intentionally seeking TVE content. The simulation was designed to ensure consistency and comparability across all platforms by the use of keyword lists, controlling for the amount of time exposed to each platform and repeating the experiments many times with different agents, thus increasing the likelihood of replicating the study's findings.

To address the six tasks defined by the study's sponsors, the following metrics were used:

- Findability: How easy is it to find harmful content, measured in the count of harmful posts during a one-hour interval?
- Removal Rate: What is the amount/percentage of harmful content removed by the platform?
- Removal Time: How much time did it take for harmful content to be removed by the platform?
- User Sentiment: What is the public (crowdsourced) perception of the appropriateness of harmful content, measured on a 5-point Likert scale?
- Amplification (filter bubble size): What is the amount of Bad Topic (TVE-related content, but not necessarily harmful content), Bad Content (TVE content), and Borderline content in a user's feed, measured as a percentage of the first 30 posts in the feed?

FINDINGS

Trust Lab's analysis found significant amounts of TVE content across all the major social media platforms as well as markets/languages. In addition, evidence of algorithmic amplification of TVE content on those platforms was also confirmed. The study identified significant differences in TVE-related performance and behaviour across all platforms and markets. The study's key findings include the following:

1. Evidence of amplification. The study validates the hypothesis that increased interaction with TVE content and Borderline content results in higher amplification of such content to users. All the platforms showed amplification of TVE-promoting content in their feeds. The percentage of TVE content in each platform's feed did not exceed 10% on average. Across all markets and platforms, the amplification of TVE content increased by 18% and Borderline content by 65%[4]. Twitter showed the highest level of amplification, YouTube ranked the second highest in amplifying TVE content, while TikTok showed the least. Regarding country and languages, platforms show the most TVE amplification occurring for Polish and German and for content related to Left-Wing political affiliation and younger Age Groups.

2. Findability Scores. Twitter had the highest findability scores of TVE content among the five platforms, and YouTube had the lowest findability score. Italian TVE content had the highest

[4] This is the average percentage change between the first and third (final) amplification measurements across all languages and platforms. See the methodology, Section 3.2, for more details.

findability score of all languages when measured across all platforms. Violent Left-Wing TVE content had a significantly higher Findability score than other TVE types.

3. Removal Rates. More than 90% of the TVE content found by Trust Lab remained on the platforms by the end of the evaluation period 8 weeks later. TikTok removed more TVE content than the other platforms; Twitter and YouTube removed the least. Arabic TVE content was removed more frequently than the other languages, while Italian TVE content was removed less.

4. Borderline Content. Repeated cycles of user interaction and evaluation suggest that the amplification of Borderline content grows over time and at a higher rate than TVE content.

5. Variability across platforms and demographic blindspots. Each platform behaves differently in recommending TVE content with similar user features and behaviours. All platforms tended to recommend TVE at increasing rates as users interacted more with it, regardless of its TVE rating. Left-Wing politically oriented recommendations were far less likely to be removed by platforms than other groups. TikTok had a significantly low rate of success in amplifying TVE content to users, while Twitter and YouTube had the highest, with the former recommending most to Right-Wing and the latter recommending TVE content even to users with low rates of interaction. Ultimately, analysis based on user demographics and platform engagement data is still inconclusive to determine the full scope of how recommender algorithms operate on the back-end; more transparency and research is needed.

TABLE OF CONTENTS

1 Objective of the Final Report

The objective of this Final Report is to describe and document the work performed during this study and the outcomes of all completed Tasks. The report is extensively supported by examples from the measurement tasks, with case studies to showcase the role and effects of the algorithmic amplification and other visual aids. The majority of sections are supported by data visualisations in the form of charts, flowcharts, tables and other forms of data representation to present the results and tell the story in the most effective way.

2 Introduction

The existence of Terrorism and Violent Extremism (TVE) content online is one of the most concerning social media trends witnessed over recent years. This trend has been exacerbated by social media algorithms that provide curated experiences to users by recommending content that keeps users engaged with the platform for longer. The severity and amount of TVE is not contained to a few small corners of the internet. It is easily accessible on major online platforms and threatens personal health and safety, peaceful co-existence, and the well-being of all citizens[5]. Social media platforms are not always able to provide a safe online experience, and policies and enforcement practices can vary substantially across platforms.

This project aims to inform the European Union Internet Forum (EUIF)'s work related to the tools and measures used by online platforms to moderate TVE content. Furthermore, this project will provide insights and analysis into the algorithmic amplification of TVE content on five leading social media platforms (Facebook, Twitter, YouTube, TikTok and Instagram) across eight markets/languages (Arabic, English, French, German, Italian, Polish, Spanish and Russian). The overarching research question is how the five social media platforms compare in exposing motivated new users[6] to TVE content. In order to answer this question, we designed a research method that allows for inter-

[5] https://cronfa.swan.ac.uk/Record/cronfa62902

[6] The term "New Users" refers to users who are intentionally looking for certain types of content.

platform comparisons and spans a broader range of platforms and languages than has typically been studied in the field of trust and safety research.

Our approach is to critically review the metrics collected relating to how social media platforms' machine learning-based content delivery systems amplify TVE content. The comparison across markets and platforms enables the ranking of individual platforms based on their performance, among other factors, to understand which platform provides the highest exposure and amplification of TVE content. Subject matter experts and industry veterans provide these recommendations to the European Commission based on assessments of the risks posed by the dissemination of TVE content and the potential for radicalisation through algorithmic amplification. Our expertise in trust and safety and online TVE and Borderline content shapes our guidance on content moderation.

3 Approach

3.1 Theoretical Framework

The goal of the study is to simulate the behaviour of motivated new users to research how social media platforms respond to users who perform targeted searches for TVE content. Motivated new users are new users on a given platform who have a targeted interest in TVE content (based on, for example, real-world interaction with TVE concepts) and who are looking specifically for that type of content on the platform.

The use of new accounts provides us with several advantages over using existing accounts. First, setting up accounts and pre-training them on 'benign' content would take a lot of time and resources that the current study scope doesn't allow for. Second, new accounts ensure that we have a more controlled research environment than we would have in the case of pre-existing accounts. In our current approach, we are aware of all the TVE-related content that agents find, and all the content that is recommended in the user feeds is captured and analysed.

If we pre-trained all the accounts on benign content, we would not only have to account for these differences in our analyses, but we would also have a harder time untangling the effects of the recommender system because of the different pre-exposure any of the accounts might have had during this pre-training phase. Therefore, starting with new accounts provides a clean baseline for comparison; even if the business models of the platforms under study are different, all our research on the platform is starting from the same entry point. In order to provide a quantitative analysis, such a common baseline is necessary.

Using new accounts also gives us the opportunity to test how quickly a recommender system will start recommending TVE content to these types of users. This information is relevant because the quicker the recommendation and amplification of TVE content occur, the more implications this has for mitigations on the platform's side. Thus, we are using the 'cold start' problem to our advantage here.

3.1.1 Addressing Replicability

Since user behaviour is at the heart of this study, the concern about the replication of findings must be addressed. In recent years, we have seen that social science research into human behaviour has

faced issues with the replication of findings[7]. This is because human behaviour is complex, and it is unwise and inaccurate to make generalisations from the results of a single study.

Behavioural studies like these will always suffer from issues with replication, at least to a degree. In order to minimise the chance of introducing unwanted variance while at the same time keeping the behaviour under study within the realm of realistic (albeit extreme/targeted), we used the following constraints:

- *Using a consistent understanding of TVE across platforms*: we have developed a set of policy definitions in accordance with the European Commission that covers a spectrum of TVE content and can be used cross-platform, enhancing the focus of our study. A consistent understanding is needed for this study, in part because platform-specific policies are inconsistent when compared to each other.

- These definitions are thus heavily informed by the existing platform definitions for TVE content but go beyond what the platforms currently define as TVE to more adequately capture the extent to which Borderline content exists on these platforms. This means that we expect content might not be removed from the platform if it is not covered by the platform's definition of TVE or that platforms might miss it during enforcement. This provides us with an opportunity to highlight the limitations of the existing platform definitions and provide mitigation strategies to address them.

Definition of TVE: Any media, including text, images, and videos, that promotes or glorifies terrorism or violent extremism or advocates for the use of violence to achieve political, ideological or religious goals.

- This category includes content that is the most explicitly promoting or supporting terrorism or violent extremism and which contains explicit calls to action or direct incitement to violence.
- The content contains explicit calls to action or statements of support for terrorism or violent extremism.
- The content contains inflammatory or provocative language that could be perceived as supportive of terrorism or violent extremism.
- The content contains images or videos that are clearly promoting or supporting terrorism or violent extremism.

❖ **Violent Right-Wing Extremism (VRWE)** are acts of individuals or groups who use, incite, threaten with, legitimise or support violence and hatred to further their political or ideological goals, motivated by ideologies based on the rejection of democratic order and values as well as of fundamental rights, and centred on exclusionary nationalism, racism, xenophobia and/or related intolerance.[8]

[7] https://arstechnica.com/science/2018/08/why-do-only-two-thirds-of-famous-social-science-results-replicate-its-complicated/

[8] The definition is provided by the DG HOME and incorporated into our methodology.

- **Violent Left-Wing Extremism** is a collective term for all efforts directed against the free democratic basic order that is based on treating the values of freedom and (social) equality as absolutes, especially as they are found in anarchist and communist ideas.

- Regulation Addressing Dissemination of **Terrorist Content Online** (EU 2021/784) (TCO) provides information for EU Member State competent authorities on reporting terrorist content. The full language of TCO can be found in the Official Journal of the European Union[9]. Under TCO, reasons for considering the material to be terrorist content are that such material:
  - Incites others to commit terrorist offences, such as by glorifying terrorist acts, by advocating the commission of such offences;
  - Solicits others to commit or to contribute to the commission of terrorist offences;
  - Provides instruction on the making or use of explosives, firearms or other weapons, or noxious or hazardous substances, or on other specific methods or techniques for the purpose of committing or contributing to the commission of terrorist offences; or
  - Constitutes a threat to commit a terrorist offence.

Definition of Borderline content: This category includes content that is not explicitly promoting or supporting terrorism and violent extremism but which may contain language or ideas that could be leading towards pathways of radicalisation. The criteria for this category include:

- content that is hard to identify as illegal or as related to violent extremism and radicalisation
- content that, despite being legal, can harm and lead to violent extremist behaviour and radicalisation (such as disinformation, conspiracy theories, which can also lead towards dehumanisation)
- tactics used to manipulate users and amplify Borderline content leading to violent extremism, such as algorithmic amplification techniques that profit from biases in content sharing algorithms.[10]

- *Having only native speakers performing searches*: the underlying reasoning is that we want to understand language nuances in the platform's approaches to amplifying and/or moderating TVE and Borderline content and to get a better understanding of potential risk areas for specific languages.

- *Structuring the search tasks when it comes to keywords and search terms*: We used a fixed set of 60 keywords, translated for each language. Keywords were derived from existing research and our own evaluation and internal testing. The 60 keywords we chose are related to common terrorism themes across all markets.

  This set is subdivided into three groups of 20 keywords for each type of TVE under study:
  - Violent Right-Wing Extremism (Local/National and Ideological)
  - Violent Left-Wing Extremism (Local/National and Ideological)
  - International TVE (Religiously motivated terrorism)

[9] Regulation (EU) 2021/784 of the European Parliament and of the European Council of 29 April 2021 on addressing the dissemination of terrorist content online, [2021], L 172/79.

[10] This definition is taken from the EUIF Handbook on Borderline Content (2023).

Personas were assigned to one of three TVE types and only used their set of 20 keywords to provide a balance between being thorough and providing comprehensive coverage. Keywords were chosen according to Trust Lab's proprietary and representative selection criteria. Within this set of 20 keywords, there were a similar number of keywords for people, organisations, hashtags and slogans (or phrases). Agents were able to select keywords from within their set randomly but were not able to add new keywords. All keyword lists were translated by native speakers into terms that are culturally relevant in the local language[11]. People's names were not translated, but hashtags and slogans may have been changed slightly in order to fit better with a specific language's nuances.

The keywords serve as a basis for the targeted searches. As such, this structure ensures that every search starts at the same place, with very few limitations with regard to how a user's journey subsequently unfolds after the initial search. Thus, we are recreating the 'rabbit hole' experience where a user might start with one video and end up exploring different videos.

The underlying reasoning for this approach is that having the same list of keywords for each language improves comparisons across platforms and languages. Some keywords will unearth more content than others in certain languages or on certain platforms, and being able to report on these differences will add a valuable component to our study. This would not be possible if we were to have different sets of keywords for each language or platform. Having different sets of keywords would also limit our ability to interpret any findings for metrics such as findability. This is because we would not be able to tell whether a certain keyword is a culprit or if a certain platform simply performs better in terms of concealing TVE-related content from its users.

[11] The list of English keywords can be found in the Appendix, Section 6.2.

3.2 Methodology

In this field of study, our methodology is comparatively unique. Instead of using bots or automated Application Programming Interface (API) queries, we use human agents (scouts) to simulate 'real' accounts. Their behaviour is designed to ensure consistency and comparability across platforms to enhance the likelihood of replication of the study's findings.

Human agent-driven investigations present certain limitations and challenges that have to be taken into consideration, such as human error, bias, and high costs associated with the method. Yet, we believe the agent-centric approach for data collection is best suited for the task at hand. One reason is that platforms' terms of service do not allow programmatic (automated) searching. Another reason is that the study aims to find out how user behaviour influences the content that is recommended to them (and, consequently, how that content influences the user), and this would not be possible with an automated approach. Lastly, and most importantly, it is the absence of primary data from the platforms themselves that place major constraints on the field of Trust & Safety research. Without access to platform internal data (access to the full corpus on content and behaviours) or a representative stratified sample (also needs to be provided by the platforms due to access restrictions) or a waiver for platform terms of service (TOS) that don't allow for programmatic access and searching, the results will be limited. In our discussions with the platforms, this point was raised multiple times, but broader data access was not provided. As a good sign for future research, the EU's Digital Services Act introduces provisions that allow access to data to researchers of key platforms and requires very large online platforms to disclose key data on the functioning of algorithms, a step very much needed to increase transparency and accountability for user safety.

The current approach is realistic yet constrained. In real life, people may not just use their accounts for searching for TVE content and might switch between different interests and topics, some of which are more innocent than others. A complete 'real life' approach would nevertheless introduce variance of a nature that would make it hard for our study to be replicable by others or to assess our results in a comprehensive manner. Therefore, we have chosen this approach while keeping these trade-offs in mind for future research.

Our study aims to understand how social media platforms' recommender systems algorithmically amplify TVE and Borderline content to users. Recommender systems, at their core, are algorithms that suggest relevant items to users - these could be products to buy, movies to watch, content to consume, and so on. Since there are many different recommender systems on different platforms, this could be an expansive endeavour. Some platforms recommend their users other accounts to follow/subscribe to, some platforms recommend content the user should consume next ("You watched this video, here's another one you might like"), and most platforms have separate 'Explore' pages that are exclusively filled with recommended content.

To ensure comparability across platforms and to keep the scope concise, we have chosen to limit this study to the recommender systems that populate the home feed of users. Feeds exist on all the platforms under study, while other recommender systems can be unique to one or several platforms and would, as such, not make as good a basis for comparison.

To compare platforms and markets directly, results need to be quantifiable, and metrics must be consistent across all platforms. To assess amplification, we have identified the following relevant metrics:

Table 3.2 - Metrics to assess amplification across platforms

Metric	Description
Findability	How many pieces of content a motivated user can find during a targeted search on the platform within a limited time window. This variable is a proxy for the amount of TVE content that is surfaceable on the platform.
Filter Bubble	A homogeneous feed/recommended content caused by the algorithmic amplification of certain content types on a given platform. The higher proportion of similar content, the deeper/larger the filter bubble.
User Sentiment	Crowdsourced severity ratings of the content surfaced. These ratings will be on a Likert scale of 1-5[12].
Removal Rate	How much content that is surfaced is being removed by the platform within a 8-week time period. This is passive removal; agents are not to report pieces of content as this would influence the algorithm.
Removal Time	How long it takes for a piece of content to be removed by the platform within an 8 weeks time period. This will be passive removal; agents are not to report pieces of content as this would influence the algorithm.

Findability is not the same as prevalence. We don't capture the total amount of content that agents come across, nor do we make any claims on how much TVE content actually exists on a platform. Findability is about how much content a person can find within a certain time frame.

Amplification in the context of our study can thus be seen as the increase in the amount of TVE content over time and the relationship between initial user searches and the amount of TVE content being recommended afterwards. Either one of these metrics indicates an amplification of TVE-related content.

In our study, we have been using two metrics ("Bad Content" and "Bad Topic") to measure the depth of filter bubbles. Bad Topic is content that is related to TVE topics but could be neutral, promoting or negative. This includes harmful TVE as well as EDSA[13] content about TVE. Bad Content is synonymous with TVE content and what is also referred to as promoting TVE. It is also a subset of Bad Topic. We capture this distinction to evaluate whether the recommender systems recommend content to users based on the overall topic of TVE versus the actual content of the posts being consumed.

For the Moderation metrics (Removal Rate and Removal Time), Trust Lab used patented technology "Kaptix", which monitors each found piece of content for removals. Kaptix is an internally developed tool used to monitor whether social media posts are live on the platform.

3.2.1 Contributing Factors to Amplifying Harmfulness

In order to better understand which factors (behavioural or contextual) contribute to amplifying potentially harmful content, we have tracked both content metadata (number of likes, shares, and comments as well as the number of followers of the posting account) and behavioural data from our agents.

[12] A graphic depiction of the values on the Likert scale is available in Figure 4.1.5 of the report.

[13] Educational, Documentary, Scientific, Artistic.

Agents were divided into High/Low interaction categories. Low Interaction agents only searched and watched content (for videos up to 5 minutes, regardless of video length), while High Interaction agents also liked and followed content. Agents did not comment or re-post/re-share the content. The influence of human interaction and contextual data on the harmfulness risk that recommender systems pose will be discussed in more detail in Task 5.

3.2.2 Process

Agents were assigned one or more personas. Within each language, 30 personas were created. Each persona had an account on each of the five platforms under study. This means that for each language, there were 10 Right-Wing TVE, 10 Left-Wing TVE and 10 International TVE agents, with 5 High Interaction and 5 Low Interaction agents per TVE Type. Agents were instructed to perform three Search Tasks and three Evaluation Tasks for each account. This allowed us to study the changes to the search results and feed recommendations over time.

All tasks were performed on mobile devices.

- **Search Task:** The Search Task started from a seeded keyword search. Agents were either in the Low or High interaction group and interacted with the content as per instructions. Agents were not allowed to add self-made keywords to searches. Agents were allowed also to browse their feeds to search for, interact with and capture TVE content according to the Interaction Type they were assigned to. Each Search Task was limited to one hour. We chose a one hour cutoff to ensure consistency among agents and due to resource constraints.

- **Evaluation Task:** The Evaluation Task started one day after the Search Task. This was to give the recommender systems enough time to update their recommendations. Agents were instructed not to interact with any content but to capture each post in their feed up to the first 30 posts.

- **Repetition:** Both Search and Evaluation Tasks happened 3 times to collect 3 data points for more accurate measurement and monitoring.

Table 3.2.2 - Number of personas per language

TVE-type \ Interaction type	Low Interaction	High Interaction
National / Violent Right-Wing Extremism (20 keywords)	5 personas	5 personas
International Terrorism (20 keywords)	5 personas	5 personas
Other / Violent Left-Wing Extremism (20 keywords)	5 personas	5 personas

With 8 languages and 30 personas per language, that means that we have 240 personas in total.

3.3 Quality

Results from the Search Tasks and Evaluation Tasks were subjected to a Quality Assurance Process to ensure that the content found and captured was indeed TVE content. For this work, we partnered with several international organisations that specialise in finding and analysing Trust and Safety-related content, as well as individuals from academia who study and work in the field of Terrorism and Violent Extremism[14].

Table 3.3 - Quality Assurance

First Level Review	Search: Agent made an assessment of what is and is not TVE-related content. The agents have had policy training to make this distinction. Specialised QA agents reviewed 100% of the content. Evaluation: The agent performs a cursory labelling exercise on the web form. 100% of this sample is reviewed by specialised QA agents.
Second Level Review	A random sample of the complete dataset was used to analyse other metrics (not Findability) to control for the quality of the labelling performed by human raters. Due to the large amount of data that was collected and complexity of the labelling task, multiple rounds of labelling were necessary by qualified individuals on a smaller, cleaner sample. Based on the specifics of the data (amount of search stages, amount of platforms and languages under study), we arrived at a maximum sample of 7200. Since not all data combinations had 30 posts, the total sample size was 7074. The data was ensured to have the following: Is the post related to TVE? (binary, based on average of rater's scale) Is the post promoting TVE? (binary, based on average of rater's scale) Is the post Borderline TVE content? (binary, based on average of rater's scale) Content Appropriateness score (scale, crowdsourced) Engagement Labelling (sourced from post data) To calculate the Findability scores, we used the complete set of content that was found during the Search phase. We relied on the first-level review to determine whether or not a piece of content was TVE.
Third Level Review	Content from the original dataset that had already been reviewed by Trust Lab experts and third-party academia experts was also included in the stratified sample as described above.

[14] More details on the consulted experts can be found in the Appendix, Section 6.3.

4 Analysis

Due to the large volume of content we captured through our work we have chosen a stratified sampling method for the analysis of our Filter Bubble metrics and Removal Metrics. Findability scores were calculated using the complete set of data that was found during the Search Stages rather than a sample. User Sentiment data was also calculated based on the stratified sample.

4.1 Task 1 - Measuring the Dissemination of TVE Content

4.1.1 Findability Main Highlights

- **Twitter has the highest Findability score of all platforms analysed**. Since we don't know how specific characteristics of platforms (such as having an open API yes/no) influence moderation of content, we do not make assumptions on the potential impact of these characteristics on the amount of content we were able to find.

- **Arabic TVE has the lowest Findability score, while Italian has the highest.** Finding TVE content through search appears to be easier in Italian, while it's harder in Arabic. There are a lot of confounding factors that could contribute to this finding, so we cannot make assumptions about the cause.

- **Left-Wing TVE has a higher Findability score than other TVE Types.** This points towards a larger trend, where International and Right-Wing TVE content are more easily identifiable as extreme content, whereas Left-Wing TVE constitutes more variations across cultures and languages.

4.1.2 Findability Findings

Our analysis shows that Twitter has the highest Findability score (2.91) on the platform[15], while YouTube has the lowest (1.71) among the platforms we analysed. This means that Facebook and Instagram, as well as TikTok, make up the middle on our scale. Findings for YouTube and Twitter are statistically significant. (See Figure A.6.4.1).

Moving to languages, Arabic TVE has the lowest Findability score (1.23), while TVE in Italian language has the highest (2.78). This means that finding TVE content through search is easier in the Italian language, while it's harder in Arabic. Both of these findings are statistically significant. (See Figure A.6.4.2).

For TVE Types, International TVE has the lowest Findability score (1.41), while Left-Wing TVE has the highest (3.01). That means that it's easier to find Left-Wing TVE, while it's harder to find International TVE by searching the platforms. These findings are statistically significant. (See Figure A.6.4.3).

Deep Dive: Violent Left-Wing Extremist Content

Since Violent Left-Wing Extremist content has a significantly higher Findability score than other TVE types, we wanted to explore this category of content in a bit more depth. The kind of content we find

[15] Examples can be found in the Appendix B, Section 6.6.

when we look for Violent Left-Wing Extremist content is mostly supportive of anti-capitalism, pro-anarchism and pro-communism[16].

These findings point towards a larger trend. Terrorist and Violent Right-Wing Extremist (TVRWE) topics, such as neo-Nazism, anti-Semitism, racism and white supremacy, frequently attract the attention of policymakers and journalists, which causes social media platforms to invest in technical and human resources to reduce the harm of such material. On the other hand, Violent Left-Wing Extremist content encompasses a range of topics that have entered modern political, social and cultural spheres without causing the same level of concern, such as anti-capitalism, general statements about the failure of modern-day institutions and the desire to learn about other forms of governance, such as communism and anarchism.

Deep Dive: User Engagement Metrics

In order to better understand how many people might have engaged with the TVE content that we surfaced, we performed an engagement analysis on four metrics: likes, comments, shares and the follower count of the accounts who posted the TVE content.

- TVE content on Instagram and YouTube is shared much less widely than content on other platforms. Content on Twitter was shared more widely. (See Figure A.6.7.1)
- TVE content on YouTube is liked more on average than other platforms. Facebook has the lowest number of likes for TVE content. (See Figure A.6.7.2)
- TVE content on YouTube is commented on more on average than other platforms. Comments on Instagram seem virtually non-existent. (See Figure A.6.7.3)
- TVE content on YouTube comes from accounts with more followers on average than on any other platform examined. (See Figure A.6.7.4)
- While YouTube has a low Findability score, the platform scores high on almost all engagement metrics, except for shares. This seems to indicate that the platform is in control of suppressing Bad Content from being found in search but that users have found other ways to access the content. (See Appendix, Section 6.7 for more details)
- Facebook and Instagram are generally at the bottom of the engagement charts. Especially Instagram scores low, having no shares and no comments on average for TVE content. (See Section A.6.7)
- Comparing the TVE vs non-TVE engagement averages shows that TVE content gets much lower engagement than non-TVE content, but these results are not statistically significant for most platforms. (See Appendix, Section A.6.7 for more details)
- It should be noted that the differences in engagement between platforms could also be due to the user interfaces of the platforms. It could be argued that YouTube and Instagram are less set up for sharing content than Twitter or TikTok, for instance. (See Appendix, Section 6.7 for more details).

4.1.3 Removal Metrics Main Insights

- **The platforms removed very few items** during the timespan we were tracking them, taking nearly a week or more to remove 50% of the content we tracked from the point we found it, resulting in TVE being shared heavily on some platforms like Facebook.

[16] Examples can be found in the Appendix, Section 6.6.2.

- **Twitter removes the least amount of TVE content**.

- **Arabic language TVE content is removed more often than other types of TVE content, while Italian TVE content is removed less**.

- **Terrorist and Violent Left-Wing Extremist content is also removed less than other types of TVE content**.

- Findability and Removal Rate are related to each other: the less likely it is that TVE content is removed, the more likely it is that we can find it. **It is, therefore, not surprising that the findings in this section resemble the findings from the Findability section.** Twitter removes the least amount of TVE content, and Arabic TVE content is actioned and removed more, while Italian TVE content is removed less, and Left-Wing TVE content is also removed less, than other types of TVE content.

- **Combining these insights seems to indicate that platforms tend to dedicate less effort to removing terrorist and Violent Left-Wing Extremist content and Violent Extremist and terrorist content in Italian language and instead focus more on removing violent Extremist and terrorist content in Arabic language, International and Right-Wing TVE content.** Since platforms still remove less than a third of the content marked as TVE Promoting, this seems to point towards inconsistent content moderation policies, in particular, less pronounced policies and enforcement focus towards markets and topics that garner less attention from the general public.

- **YouTube scores low on Removal Rate, which is an interesting reversal case:** The platform does comparatively well with suppressing TVE content in search without removing such content from the platform. See Section 4.2.2 for a deep dive into this.

4.1.4 Removal Metrics Findings

Our analysis shows that TikTok removed more TVE content than other platforms (11% of the content removed), while YouTube removed the least amount of content (3% of the content removed). Overall, platforms removed less than a third of the content marked as TVE promoting. These findings are statistically significant. (See Figure A.6.8.1).

Arabic language TVE content is more often removed than most other languages (10% of the content removed), while Italian language TVE content is less often removed than most other languages (3% of the content removed). These findings are statistically significant. (See Figure A.6.8.2).

TVLW content is less often removed than other TVE types: only 4% of the time, compared to 7% (Right-Wing TVE) and 8% (International TVE). This finding is statistically significant. (See Figure A.6.8.3).

The platforms removed very few items during the timespan we were tracking them, taking nearly a week or more to remove 50% of the content we tracked from the point we found it. All Removal Time metrics are anecdotal because of the small number of removals overall. At each interval, the

percentage represents the amount of content that was removed out of the total removed within an 8-week period. (See Figure A.6.9.1).

Generally, German language TVE content was removed more quickly, while Italian TVE content was removed more slowly than other languages we've studied. (See Figure A.6.9.2) TVLW content also seemed to be removed the slowest, while International TVE content seemed to be removed more quickly. (See Figure A.6.9.3).

Deep Dive: Removal Metrics

The low rates of removal by themselves do not tell a very satisfactory story, so a deeper analysis was conducted to look into the impact that this is causing based on the data that was collected. The focus of this analysis was to contrast removals with user engagement and to review in particular, the number of shares. Shares were chosen because this is a much more deliberate action that also highly impacts the dissemination of TVE content than other forms of engagement[17]. It's worth noting that shares are most likely the least accurate for platforms like YouTube that have a significant presence on desktops because sharing can be just as easily done by copying the URL from the address bar. Also, sharing off-platform (for example, sharing a link on another social network) may not be trackable by the content owner.

Of the content that was eventually removed from platforms within the study period, TikTok TVE content was disseminated the most, with 120 shares per TVE post on average (see Figure A.6.8.4). However, of the TVE that wasn't removed by platforms Facebook has the highest number of shares per post 341 followed by TikTok with 283 on average (see Figure A.6.8.5).

The proportionally higher removal rates of TikTok contrasted with higher average number of shares suggests that TikTok TVE content is being consumed and spread at higher rates than other platforms. Thus, despite TikTok's higher performance on other metrics, it is struggling to keep up with its user base and high throughput even on high-harm content like TVE. This same trend appears in Borderline content except with higher shares counts (see Figure A.6.8.7).

Looking at how different languages impact TVE content that wasn't removed, French and Italian have more shares on Facebook while other languages are more prevalent on TikTok (see Figure A.6.8.6). Borderline content in German language, however, sees more shares on Facebook which is the only language that changes between TVE and Borderline content (see Figure A.6.8.9).

[17] Ljungberg, J. et al, Like, Share and Follow: A Conceptualisation of Social Buttons on the Web, July 2017.

4.1.5 User Sentiment Main Insights

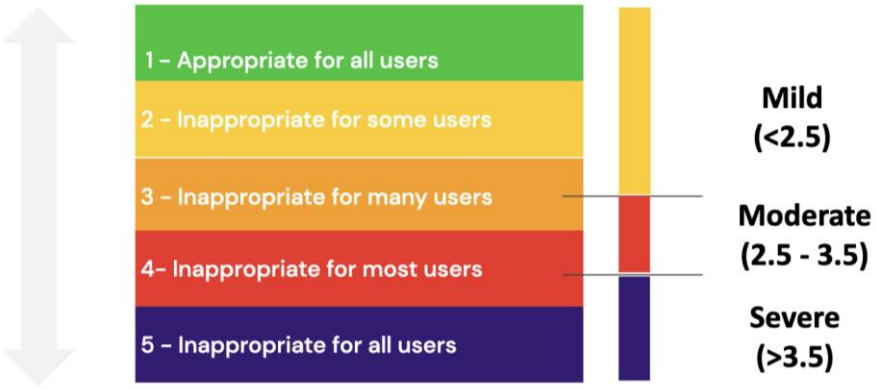

Figure 4.1.5 - Appropriateness score

- **Platforms are mostly similar** in terms of **severity of content based on user ratings**, indicating that severe content is spread roughly equally across the platforms. Italian content is rated as the most severe, while German is rated as the least severe. Most differences between languages and their severity ratings are statistically significant.

- **TVE content in Italian is rated as the most severe, continuing a trend that seems to indicate that Italian language TVE content is largely going unnoticed by platforms**. Other languages are also mostly statistically significantly different from each other, indicating that users of different languages have potential cultural differences when it comes to how harmful they judge TVE content to be.

4.1.6 User Sentiment Findings

Platforms are mostly similar in terms of how severe users rate their content, indicating that severe content is spread roughly equally across the platforms. (See Figure A.6.10.1).

Italian language TVE content is rated as the most severe (98% rated as severe), while German language TVE content is rated as least severe (43% rated as severe). Most differences between languages and their severity ratings are statistically significant. (See Figure A.6.10.2).

Right-Wing TVE content is viewed as less severe by users than other TVE types (78% rated as severe), which is a statistically significant finding. This is interesting and could be due to the fact that users are getting more used to seeing Right-Wing TVE content, but also to the fact that platforms are removing the more severe Right-Wing TVE content from their platforms more readily. (See Figure A.6.10.3).

Severity Over Time

Does searching for and interacting with TVE content influence the severity of the content that is subsequently found and recommended to users? A trend here could be an indication of amplification, not necessarily in the *amount* of content that a user sees, but in the *harmfulness risk* of the content that users engage with.

No clear trends emerge when looking at the severity ratings over time per platform. (See Figure A.6.10.4).

4.2 Task 2 – Measuring the Role and Effects of Automated Dissemination of TVE Content

4.2.1 Filter Bubble Main Insights

- **Taking all platforms, languages and TVE types together, there is an amplification of all content types over time**. This amplification is only significant for Bad Topic and Borderline content, but not for Bad Content.
- **All platforms show signs of amplification after the first Search phase, showing Bad Content in the user's home feeds, but platforms do not show any significant further amplification of Bad Content over time.** On average, no more than 10% of any feed consisted of Bad Content.
- **All languages show signs of amplification after the first Search phase, but only Italian shows significant further amplification of Bad Content over time.** French does show a significant further amplification of Borderline and Bad Topic content over time.
- **All TVE Types show signs of amplification after the first Search phase, but none show significant further amplification of Bad Content over time.**
- Overall, these findings seem to indicate that although there seems to be an amplification of Bad Content in the feed after showing initial interest by a user, this is not significantly further amplified when the user continues to engage with said content.

4.2.2 Filter Bubble Findings

Taking all platforms, languages and TVE Types together, there is an amplification of all content types over time. For Bad Content, the main focus of our study, our amplification metric increased from 4.4% to 5.2%, an increase of 17.5%. Looking at Bad Topic, which includes TVE and non-harmful content related to TVE topics, we saw an increase from 6.9% to 9.6%, which is statistically significant (+39.1%). Additionally, we looked at Borderline Content, which amplified from 3.3% to 5.4% (+65.1%), which is also statistically significant. (See Figures A.6.11.1).

Instagram is the only platform that shows a significant amplification across the three stages for Bad Topic content (+118%), as well as Borderline content (+150%). None of the platforms show a significant further amount of amplification for Bad Content, although Facebook doubled the amount of recommended Bad Content between the First and Third stage. TikTok showed no increase, and YouTube had a negative amplification (-23%): all not statistically significant. (See Figures A. 6.11.8a-e).

Italian and French are the only languages that show significant amplification of TVE related content over time. Italian is significant for Bad Content only (+365%). The French amplification is only significant for Bad Topic (+293%) and Borderline content (+685%), not for Bad Content. Polish has the highest average percentage of Bad Content in the feed (11%), Arabic has the lowest (2%). (See Figures A.6.11.9a-h).

None of the TVE types show significant amplification over time. (See Figures A.6.11.10a-c).

Higher interaction with TVE content does result in a higher initial percentage of Bad Content in the feed (8% for High Interaction, 1% for Low Interaction). (See Figure A.6.11.12).

Deep Dive: How Platforms Reduce TVE Content

When looking across the Findability and Filter Bubble metrics, we can compare how platforms choose to reduce the spread of TVE content. (see Figure A.6.11.13).

- **YouTube:** has the lowest Findability score but a high average percentage of Bad Content in the feed.
- **TikTok:** reflects the opposite trend, having a moderate Findability score and the lowest average percentage of Bad Content in the feed.
- **Twitter:** has the highest Findability score and also the highest average percentage of Bad Content in the feed.
- **Instagram:** has a moderate Findability score paired with a moderate average percentage of Bad Content in the feed.
- **Facebook:** has a moderate Findability score paired with a moderate average percentage of Bad Content in the feed.

YouTube seems to have tighter filtering of search results, while TikTok has tighter filtering of the recommendations in the feed. Relatively speaking, Twitter has the most on both dimensions. Facebook and Instagram appear to have moderate amounts of filtering in both search and feed.

These findings could be the result of YouTube's and Twitter's algorithms having stronger personalisation, combined with a lack of filtering. When platforms strongly personalise content based on past activity but lack adequate filtering of Bad Content, the filter bubble effect becomes most problematic. TikTok's algorithm is known to have strong personalisation effects[18], but seems to have better filtering of TVE-related content than YouTube and Twitter.

4.2.3 Borderline Content

Special interest from the European Commission has been expressed with regard to the effect of the spread of certain types of Borderline content, such as disinformation and some forms of hate speech that can lead to radicalisation. This section will shine more light on how Borderline content plays into the analysis of the metrics that we calculated. It is content that does not meet the thresholds to be labelled as terrorist and violent extremist content, but that can still lead to violent extremism and radicalisation pathways.

Amplification of Borderline Content

- Overall, Borderline content has the lowest initial amplification of all TVE-related content types, but the percentage of Borderline content that is recommended to users has the highest amplification over time of all TVE-related content types.
- Instagram and Twitter are the only individual platforms that show a significant amplification in the amount of Borderline content that is recommended to users over time. Instagram shows 3 times more Borderline content when comparing the First to the Third phase.
- French is the only language that shows a significant amplification of the amount of Borderline content being recommended to users over time. 8 times more content is recommended to users in the Third phase compared to the First phase.
- None of the TVE Types shows a significant amount of amplification for Borderline content.

[18] https://www.theguardian.com/technology/2022/oct/23/tiktok-rise-algorithm-popularity

(See Figures A.6.11.8a-A.6.11.10c).

Removal of Borderline Content

- Overall, all platforms act similarly when removing searched or recommended Borderline content from their platforms. TikTok removed more searched content than YouTube and Twitter but did not remove any recommended content at all. Since we are comparing all platforms based on the same definition of Borderline content, this gives us a clear picture of which platforms are allowing more Borderline content on their feeds.
- Arabic Borderline content is removed more often during Search than most other platforms. None of the values for Evaluation are significant.
- Borderline content that could lead to Violent Left-Wing Extremism is removed less often during Search than other TVE Types. None of the values for Evaluation are significant.
- Removal Times for all TVE content and Borderline content are largely similar for all dimensions.

(See Section 6.12 for figures).

4.3 Task 3 - Assess the Risk posed by the Automated Dissemination

4.3.1 Risk Assessment Main Insights

- **Amplification: What is the prevalence of TVE content in the platform's feed?** Twitter was the platform with the highest level of amplification of Bad Content to the user, followed by YouTube, Instagram and Facebook. TikTok had the lowest level of TVE content amplification or harmfulness.

- **Risk Ratings.** We assigned a composite risk rating for each platform, based on an evaluation of three essential components: Findability, Amplification and User Sentiment. The risk ratings for the platforms mostly followed the percentage of TVE content found in the feed because the feed is where the bulk of the TVE content consumption likely occurs. The risk rating was consistent with removal and engagement rates, and there were fairly big differences among the platforms on this risk metric. Twitter had the highest risk rating of the platforms; TikTok had the lowest risk rating. Facebook, Instagram and YouTube had medium risk ratings.

- The average percentage of TVE content in the feed did not exceed 10%, even for motivated users, and did not seem to increase above 10%, even under repeated user interactions with TVE content. If platforms were only focusing on personalisation, this percentage would likely have been higher. This suggests that platforms either do not personalise much in general or that platforms limit personalisation, when TVE content, topic or Borderline content is detected. That said, when we compared the platforms on different dimensions, we observed clear differences among the platforms and further room for improvement, particularly for platforms such as Twitter with higher risk.

Recommender systems (RSs) are increasingly being used to support decision-making and improve user experiences across a wide range of industries and applications. However, it is important to acknowledge and effectively manage both technical and non-technical risks in order to reap their benefits and promote user safety fully. As requested by the European Commission, this report provides a comprehensive overview of recommender systems and the technical and non-technical risks associated with recommender system algorithms and outlines a range of mitigation strategies to help reduce risks and ensure best practices for the recommender systems and those who use them.

4.3.2 Content Moderation Background

Content moderation, or the process of monitoring, reviewing, and controlling user-generated content on online platforms to ensure compliance with community guidelines, terms of service, or content policies, is a fundamental part of the internet's infrastructure and success. Content moderation involves identifying and removing inappropriate, harmful, or offensive content, as well as addressing violations of rules and regulations set by the platform or applicable laws. The way a platform is designed and how users interact with it significantly influence the posting and interaction of user-generated content, consequently impacting how companies perform content moderation on their services (Grimmelman, 2015).

As discussed, platforms which rely on user-generated content often utilise recommender mechanics to deliver personalised content to users, deliberately surfacing tailored content most likely to keep the user engaged (O'Callaghan et al., 2014). Recommender systems have the potential to inadvertently recommend terrorist or extremist content to users who are not actively seeking such

material. This occurs because the algorithms prioritise user engagement without parallel automated objectionable content detection, allowing such terrorist or extremist content to be surfaced based on predicted engagement levels. Recent studies indicate that inadvertently exposing users to extremist content based on their expressed interest can lead to a reinforcement of extremist content in their feed; a consistent consumption of extremist content may contribute to a shift in their worldview (Edwards & Gribbon, 2013). Platforms face a challenging task in distinguishing between intentional bad actors and users who unintentionally engage in abusive behaviour, adding complexity to the moderation ecosystem. The following breakdown of prevalent filtering mechanisms and technological content moderation systems highlights the intricate balance between preserving the rights of users to freedom of expression and the challenging task of detecting and removing content in a scalable manner.

Content-Based Recommendation Systems

Content-based filtering is a technique used by online platforms, including user-generated content platforms like YouTube, to analyse a user's preferences and create a content engagement profile based on keywords and tags. This approach considers factors such as titles, descriptions, and tags to determine similarities and relevance when recommending videos, particularly for new or unseen content (Zhang, Lu & Jin, 2021). While content-based filtering offers scalability and independence from other users' data, it has limitations in the type of content it suggests to users. In the context of moderating terrorist content, platforms that rely heavily on content-based filtering pose challenges in effectively identifying and addressing such material. The focus on matching user preferences may overlook objectionable content that falls within the grey area of extremism. As the content becomes more extreme, the videos the system relies on for tags may begin to overlap between less extreme and more extreme content, making it less likely for a target audience to report such material to the platform. This increases the risks associated with content moderation and the potential for terrorist or extremist content to go undetected or unaddressed.

Although platforms basically work with some combination of recommender and flagging sub-systems, these can be brought together in different ways, sometimes without a clear demarcation between the two. Also, while these subsystems draw upon the basic recommendation algorithms and flagging approaches outlined here to optimise for the metrics specified in section 4.4.2, there are many variations of these algorithms and metrics. Also, the practical pressures of running a business (e.g., generating revenue and retaining users, keeping infrastructure and machine costs low as well as addressing user and customer complaints and escalations) implies that these systems have accumulated numerous optimisations as well as business rules, making them exceedingly complex. As a real-world example, please refer to the recently outsourced Twitter recommender system.

Automated Flagging System

Automated detection systems utilise various techniques such as machine learning algorithms, near-neighbour algorithms, fuzzy hashing models, and MD-5 hashing models to identify harmful or platform-violating content, including both well-known and unknown terrorist and extremist material. These systems bring potentially illegal or policy-violating content to the attention of a team of content moderators, either directly employed by the platform or affiliated with a vendor company. Depending on the specific model employed, there are cases where reports are automatically closed if signals suggest the content is spam, does not violate the platform's Terms of Service, or has been previously reviewed by a content moderator. Automated detection systems play a crucial role in helping online platforms identify more severe content, initiating the cycle of moderation and policy enforcement on the platform.

User-Based Flagging System

User-based reporting systems rely on users to flag content that they believe violates an online platform's policies or is deemed objectionable. These reporting systems vary across platforms, often employing dedicated reporting forms or allowing users to directly flag content they find problematic while viewing it. User flags serve as a means for users to report content that may not have been detected by technical solutions, giving them a voice in the content moderation process. After users submit reports, the platform's team of content moderators reviews them and determines the appropriate actions to take regarding the reported content or offending user account. Many large online platforms deprioritise these user reports in favour of automated detection, but will no longer be allowed to do so under the new European Union's Digital Services Act, which places emphasis on users' right to report and appeal content they find objectionable, including terrorist and extremist content.

4.3.3 Risks Assessment

How can platform algorithms cause filter bubbles?

The term "filter bubble" is generally credited to internet activist Eli Pariser circa 2010. In Pariser's influential book, The Filter Bubble (2011), it was predicted that individualised personalisation by algorithmic filtering would lead to intellectual isolation and social fragmentation. To increase user engagement, social media companies may connect users with ideas they are already likely to agree with, thus creating echo chambers of users with very similar beliefs. The concern is that recommender systems may influence users to engage in progressively narrower content domains and in directions they might not otherwise have pursued. The EU Counter-Terrorism Coordinator has argued that the amplification of legal but harmful content may be conducive to radicalisation and violence because it normalises it and exacerbates polarisation in society.

The Global Partnership on Artificial Intelligence (GPAI) has also identified that recommender systems can amplify user bias related to TVE content, as follows:

> The key issue for recommender systems . . . is that social media users are known to show small biases towards extreme content of various kinds, that act as another influence on the content items they engage with. For instance, they have a tendency to share political messages containing 'moral emotional expressions' (Brady et al., 2017; Brady and van Bavel, 2021), particularly negative ones (Crockett, 2017; Brady and van Bavel, 2021), messages that refer to a political 'out-group' (Rajthe et al., 2021), and messages that contain falsehoods (Vosoughi et al., 2018). If these biases persist while a user interacts with a recommender system, the system's repeated updates of its user model may lead the user towards messages containing increasing levels of negative political emotions, an increasing focus on political out-groups, and increasing amounts of misinformation – and potentially towards domains of violent extremism. Again, our earlier report (GPAI, 2021) presents these concerns and the studies that support them in detail.[19]

[19] GPAI 2022, *Transparency Mechanisms for Social Media Recommender Algorithms: From Proposals to Action. Tracking GPAI's Proposed Fact Finding Study in This Year's Regulatory Discussions.* Report, November 2022. Global Partnership on AI.

Recently, the Global Internet Forum to Counter Terrorism commissioned a survey of the existing empirical literature on the issue of whether recommender systems promote extremist content and whether online users are "radicalising by algorithm".[20] In the literature review, 10 of 15 studies demonstrated that recommender systems can promote extreme content.

Can bad actors exploit platform algorithms?

Terrorists could "game" flagging systems by experimenting and learning over time how to have their content bypass or circumvent either automated or manual flagging systems. According to certain researchers, the survival of certain TVE content on platforms relates to the way in which terrorist organisations have learned to modify their content to evade platform controls. Tactics include: breaking up text and using strange punctuation to evade platform tools that would search for keywords; blurring the terrorist organisation's branding or adding the platform's own video effects; mixing the terrorist organisation's material with content from real news outlets; adding the branding of mainstream news outlets over the top of the terrorist organisation's content; hijacking platform accounts; and posting tutorial videos to teach other terrorists how to do the same.[21]

Terrorists could also employ fake engagement and fake profiles to promote their content, and if undetected by platforms, recommender systems could unwittingly further promote such content. From gaming of views[22] to crowdturfing and astroturfing[23], there are a variety of technical approaches to gain (perceived) popularity on social media, which are already known to be leveraged by foreign actors to influence operations. For instance, in one study, the BBC has reported that a network of Facebook groups was used in an attempt to change perceptions related to the ongoing war in Ukraine.[24]

In this section, we presented many possible risks that can cause the promotion of TVE content on platforms. While we do not have access to internal platform documentation to understand how each individual platform's algorithms and moderation systems work, we can assess risk based on externally observable metrics from this study. We do this in the next section.

4.3.4 Risk Metrics

The previous sections described the various risks that recommender systems present, but without access to each platform's internal, proprietary algorithms and data, we were unable to quantify platform risk directly based on these factors. Instead, we evaluated the overall risk of each platform based on the data collected for this project, which could be observed outside-in. We considered three factors:

---

[20] Whittaker, J., Recommendation Systems and Extremism: What Do We Know? – Global Network on Extremism and Technology, Insights, 17 August 2022. https://gent-research.org/category/insights/).
[21] Corera, G., *ISIS 'still evading detection on Facebook', report says,* BBC News, 13 July 2020.
See also Nimmo, B. and Hutchins, E., *Phase-based Tactical Analysis of Online Operations (The Online Operations Kill Chain: A model to analyze, describe, compare, and disrupt threat activity from influence operations to cybercrime), Carnegie Endowment for International Peace, March 16, 2023.*
[22] The Flourishing Business of Fake YouTube Views.
https://www.nytimes.com/interactive/2018/08/11/technology/youtube-fake-view-sellers.html
[23] https://en.wikipedia.org/wiki/Astroturfing
[24] Putin's mysterious Facebook 'superfans' on a mission. https://www.bbc.com/news/blogs-trending-61012398

- User Sentiment: This is a proxy for the severity of TVE content found
- Findability: This is a proxy for the amount of TVE content that is surfaceable on the platforms
- Amplification: This is a proxy for the prevalence of TVE content in the feed.

We saw relatively minor differences in User Sentiment (see Figure A.6.10.1) among platforms. The platform with the highest level of Findability of TVE content (see Figure A.6.4.1) was Twitter, followed by Facebook, Instagram, and TikTok. The platform with the lowest level of Findability was YouTube.

Risk Outcome

We have assigned a composite risk rating for each platform based on our evaluation of three component metrics: User Sentiment, Findability, and Amplification. The risk ratings below incorporate information from the TVE content found in the feed and search and is consistent with removal and engagement rates.

Table 4.4.4 - Risk Rating per Platforms

Twitter	Youtube	Instagram	Facebook	TikTok
Highest	Medium	Medium	Medium	Lowest

The data we collected is based on simulating motivated users looking for TVE content and does not represent the experience of an average user of the platform. We also did not have access to internal proprietary platform data, which would have been needed for comprehensive measurement. That said, as previously discussed in the introductory section, it made sense to simulate and study the motivated user scenario. Also, the consistency observed for different metrics, such as feed prevalence, removal rates and engagement rates, suggests that the data is not an outlier.

The next section presents possible mitigation measures to consider, both based on observed data as well as conceptual risk factors discussed in previous sections.

4.3.5 Mitigation Strategies

There are a few different ways in which platforms can counter filter bubbles:

- By effectively detecting terrorist and extremist content, platforms can proactively filter such content from user feeds and search results. The **Global Internet Forum to Counter Terrorism** (**GIFCT**) is an Internet industry initiative to share proprietary information and technology for automated content moderation. As noted on the GIFCT Wikipedia page, GIFCT maintains a database of perceptual hashes of terrorism-related videos and images that are submitted by its members and which other members can voluntarily use to block the same material on their platforms. The material indexed includes images, videos and will be expanded to include URLs and textual data such as manifestos and other documents.

- Filter bubbles and echo chambers can be countered by having recommender systems explicitly optimise for content or opinion diversity. There is evidence suggesting that personalisation algorithms can enable exposure to "ideologically cross-cutting content" from outside self-selected echo chambers. (Flaxman et al., 2016). It has also been suggested that they can also be leveraged in improving the automated detection of radicalisation and for facilitating targeted counter-messaging interventions to combat radicalisation. (Schmitt et al., 2018).[25] Recent literature has also suggested that algorithms can be the cure for algorithmic filter bubbles. [26]

- Platforms should ensure that the algorithms themselves don't have ideological bias. A recent Twitter study ("the most comprehensive audit of an algorithmic recommender system and its effects on political content") indicated that the political right enjoys higher amplification compared to the political left. Our own data presented in previous sections also examines differences between Right- and Left-Wing TVE content both in terms of discoverability and moderation. Armed with such data, platforms can take a deeper look at their own algorithms and implement necessary mitigations.

[25] Wolfowicz, M., Weisburd, D. and Hasisi, B. (2021), Examining the Interactive Effects of the Filter Bubble and the Echo Chamber on Radicalization , Journal of Experimental Criminology (2023) 19:119-141 at 136. https://doi.org/10.1007/s11292-021-09471-0

[26] Gao, C. et al., Counterfactual Interactive Recommender System (CIRS): Bursting Filter Bubbles by Counterfactual Interactive Recommender System, Computing Research Repository (CoRR) in arXiv, abs/2204.01266 (2022).

4.4 Task 4 - Compare the Phenomenon Between Platforms Sharing the Same Business Model

4.4.1 Comparison Between Platforms Main Insights

- Among the platforms, Twitter surfaced and recommended the most TVE content to its users and performed below other platforms in terms of TVE content removal.

- Although YouTube's Findability score was the lowest among the platforms, its removal metrics suggested it has room to improve on removing TVE content that exists on the platform. YouTube generally scored higher than the other platforms on user engagement (e.g., likes, shares, comments, number of followers) with TVE content. This seems to indicate that YouTube was effective in suppressing TVE content from users in search but still enabled users to find TVE content through other means and interactions.

- TikTok performed moderately when it came to the Findability of TVE content, but of all the platforms, it recommended the least amount of TVE content to users. This suggests that TikTok may filter its feed more tightly than its search, resulting in no significant amplification of TVE content in the feed. TikTok was generally slow to remove TVE content; TikTok took more than four weeks to remove the content we identified as TVE. TVE content that the platform didn't remove was shared twice as much as TVE content that the platform did remove.

4.4.2 Platforms

Of the platforms, Twitter surfaced and recommended the most TVE content to its users. Twitter was lower than other platforms when it came to TVE content removal. We note that this study was conducted during a period when significant governance, policy and safety process changes were occurring at Twitter. As such, Twitter's results may be significantly different compared to earlier in the year.

YouTube has come under scrutiny during other research projects[27] and has since made improvements. Although YouTube's Findability score is the lowest among the assessed platforms, its Removal metrics combined with YouTube's engagement metrics for TVE content suggest the platform still has room to improve on removing TVE content that already exists on the platform.

In our study, YouTube recommended the most TVE content to its users; Twitter was second. It should be noted, however, that all the platforms had weak Removal rating scores, removing 10% or less of the content we marked as TVE-related.

YouTube generally scored higher than the other platforms on user engagement with TVE content. This seems to indicate that YouTube is effective in suppressing TVE content from users in search. But, it still enables users to find TVE content through other means and interactions.

[27] https://www.technologyreview.com/2020/01/29/276000/a-study-of-youtube-comments-shows-how-its-turning-people-onto-the-alt-right/
https://policyreview.info/articles/analysis/recommender-systems-and-amplification-extremist-content

4.4.3 Languages

Italian TVE content is an interesting case with regard to many of the metrics considered. The Findability score for Italian TVE content was one of the highest, and removal metric scores were in the lowest category. Under the study, Italian users in the sample were also the ones who rated their TVE content as the most severe of all the languages. This suggests that Italy may not be a high-priority market for platforms in terms of Italian-specific automated detection methods, Italian moderators, or other measures integral to a Trust and Safety enterprise.

TVE content in Arabic has long generated special attention from the platforms' moderation teams. Its Findability score is the lowest. Significantly, Arabic TVE content, especially International Arabic TVE content, is more often removed than content from other languages. Since the Paris attacks in 2015, there has been more focus from policymakers and media outlets on social media platforms to police the spread of International TVE from Arabic-speaking countries and regions, which could explain this finding.

4.4.4 TVE Type

Left-Wing TVE content treatment stands out. It's easier to find Left-Wing TVE content on the different platforms, especially in the Italian language. Left-Wing TVE content is also the least often removed. Apparently, this type of TVE content is also not as readily on the radar of platforms, possibly because Left-Wing TVE content covers a broad spectrum of content that is not TVE related per se.

This can be contrasted with Right-Wing and International TVE content, which receives more attention from policymakers and the media, thus incentivising social media platforms to create policies to moderate this type of content more readily.

Overall, one conclusion may be that the public attention to certain types of TVE content, often related to current events, plays a large role in what content gets moderated on platforms. If this is true, one could argue that this is an insufficient way of moderating TVE-related social media content. The platforms could certainly do more to be more consistent in their moderation practices (see also Task 3).

4.4.5 Borderline Content

When we look across the entire dataset, Borderline content appears to be amplified the most over time out of all the TVE-related content types that we've studied, even though Borderline content has the lowest amount of initial amplification.

Zooming in on different dimensions, such as platform, language, or TVE type, this trend is less pronounced. For example, Borderline content does not appear to be significantly amplified over time on any platform except for Instagram and Twitter. French is the only language where Borderline content was significantly amplified over time. Borderline content, potentially leading to terrorist and Violent Right-Wing Extremism, was not significantly amplified over time.

Platforms act similarly when removing searched or recommended Borderline content from their platforms. TikTok removed more searched content than YouTube and Twitter but did not remove any

recommended Borderline content. Borderline content in the Arabic language is removed more often during Search than other platforms, but none of the values for Evaluation are significant.

Borderline content potentially leading to Violent Left-Wing Extremism is removed less often during Search than other TVE Types. None of the values for Evaluation are significant. These findings are similar to other TVE-related content, which seems to show that the policies of the platforms do not place special focus on Borderline content.

4.5 Task 5 - Assessing the Risk of Radicalisation due to Algorithmic Amplification

Assess the level of risk of radicalisation to violent extremist ideologies and terrorism linked to the voluntary or involuntary automated dissemination of terrorist and violent extremist content through the use of machine learning-based algorithms or less sophisticated techniques.

4.5.1 Introduction

We employed various methods to analyse the data from the experiment and build models to predict whether a given social media platform user's demographics and behaviour could relate to TVE content amplification. In this section, we explain our process of analysis, and key results and briefly conclude with some final insights and recommendations.

Before proceeding, we note that machine learning processes are by nature exploratory (e.g., from selecting an appropriate way to organise the data, to setting up prediction problems and of course, to the selection of methods to use and the final analyses). Given the complexity of the data – and the problem – there was, naturally, significant exploration along the "machine learning pipeline" (from data structuring to final insights). We only discuss the processes, methods and results we found to be most appropriate.

While there are consistent observations across the different approaches discussed, each also provides some specific insights. Overall, consistently across all approaches outlined below, the results indicate that there is some predictability about whether TVE content is presented to users (during the Observation phase – see below).

As described in the previous sections, the experiment was carried out by having real agents impersonate generated user personas with different demographics (age, gender, country of origin and language), extreme political orientations (Violent Right-Wing Extremism, Violent Left-Wing Extremism, or apolitical/international extremism), and behaviours with TVE content (high interaction or low interaction). Over the course of three separate sessions (Search Stages 1, 2, and 3), users logged in and directly searched for and engaged with TVE content (the Interaction Phase) and, on a separate occasion, would log in again to observe the resulting content presented/recommended to them by a platform (the Observation Phase). The TVE content discovered in either phase was recorded (such as date, description, URL, etc.), its relation to TVE was rated by multiple experts (Borderline, promoting, or relating to TVE), and its popularity noted (account followers and the comments, likes, and shares for a given post). The raw data was organised into rows, with each row representing a particular post that was either found during the interaction or recommended during observation by a given persona on a given platform during a given session/login.

We study the effects that demographics, deliberate TVE searches, and the temporal aspect of doing so over multiple instances have on the platform algorithm's tendency to amplify such content. The process outlined below involves a preliminary exploratory analysis of the data using statistical descriptions and clustering by recorded content posts, followed by predictive modelling for TVE presence by session (login) for the Observation sessions – while using data also from the interaction sessions. Of course, if there is no (or weak) "signal" in the data, for example, for some platforms, one cannot make inferences reliably. Overall, the data did prove to be challenging to analyse.

Data Preparation

Data Cleaning

The data from the experiment's implementation contained very few instances of missing data. For example, only 54 (0.76%) out of 7074 rows (selected content posts across all sessions) used for the analysis contained discrepancies that rendered them unusable. This small fraction allowed us to directly remove those observations without hindering the quality of the analysis. Different features included in the data set were useful for different sections of this report, but overall the features kept for the analysis were the aforementioned demographic, behavioural and political data on the personas and the popularity and TVE ratings of the content they either found (Interaction phase) or were presented/recommended (Observation phase).

Encoding

Encoding is a process of converting data from one format to another. For example, it is used to convert categorical data (like gender or country) into numerical data so that it can be used in statistical models. There are many different types of encoding techniques, but two of the most common ones are ordinal encoding and one-hot encoding.

- *Ordinal encoding* is a method of encoding categorical data where each category is assigned a unique numerical value. This encoding is especially useful for including information about the relative significance or order of categories with respect to each other.
- *One-hot encoding* is a method of encoding categorical data where each category is represented by a binary column in the dataset. For example, if we use the Gender feature, 'male' could be labelled as {0,1} while 'female' as {1,0}. This technique is useful when there is no natural order or hierarchy among the categories, and each category is equally important.

The appropriate encoding method to use depends on the type of data. As we progressed through the analysis, we used different forms of encoding depending on the method implemented, which is mentioned in its relative section below.

4.5.2 Analysis of Content Posts

Statistical Description

We first performed exploratory analyses to determine the significance of each data feature collected and get an overview of each platform's frequency of TVE content dissemination. Once we have an overview description of the data, we can move to more complex analyses.

We start with statistics, also considering the temporal nature of the three cumulative search stages. We first map the occurrence of TVE-scoring content (Borderlines, Promotes, and Relates to TVE) based on the given Search Stage, Platform, and Interaction Level during the Observation Phase, meaning the content that was recommended to the persona after a period of engagement (*Figure 4.5.2a*).

This initial analysis indicates the following observations:

- There is overall (but different across platforms) a higher rate of TVE in the final third observation stage than the first two stages.
- There is a significant difference in the observed TVE content for Low-Interaction personas versus High-Interaction, but less so for YouTube.

- TikTok appears to have overall very low TVE content, regardless of one's Interaction Level, with almost no instances of TVE content being present in the Observation Phase.

Finally, we also noted from the coefficients of all three TVE score categories that Relates to TVE was an appropriate category that encapsulates the average score of Borderlines and Promotes TVE. We, therefore, focus on that metric.

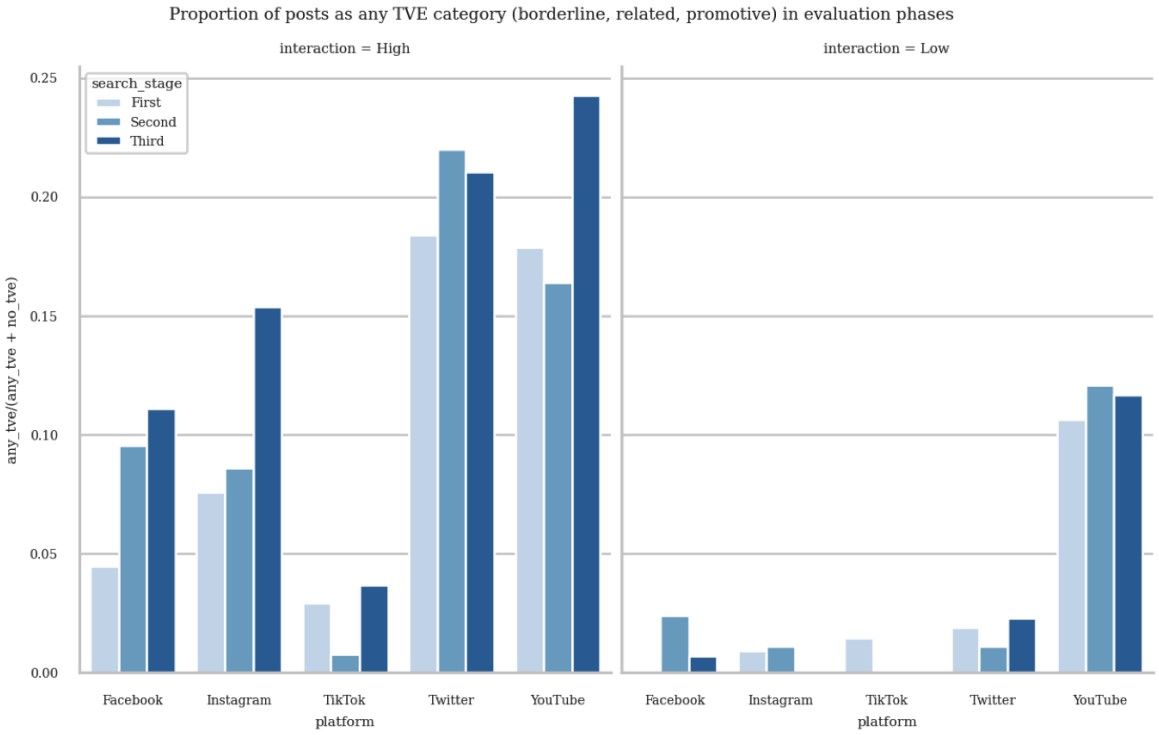

Figure 4.5.2a - Prevalence of all TVE categories, grouped by platform and divided between high and low levels of persona interaction

To further explore potential relationships, we plotted two versions of groupings of our significant demographic features and the TVE score. These features' importance is shown below (*Figure 4.5.2b; Figure 4.5.2c*).

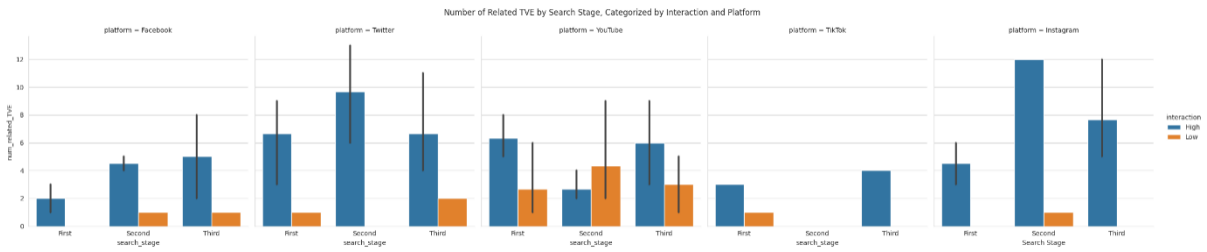

Figure 4.5.2b - Prevalence of related TVE across platforms, grouped by level of persona's interaction per search stage

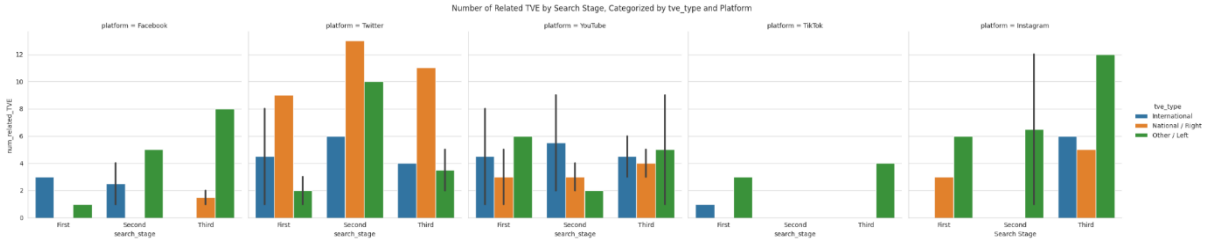

Figure 4.5.2c - Prevalence of related TVE across platforms, grouped by persona's political orientation per search stage

The results of these graphs display similar observations as the first statistical description (*Figure 4.5.2a*), with TVE presence generally increasing as the sessions accumulate. Twitter seems to cater most to recommending Right-Wing users TVE content while (again) TikTok has little to no instances of TVE content. Three observations can be made about the different platforms:

- YouTube's TVE content does not depend as much on one's interaction level, as above, or political orientation.
- As above, TikTok has overall little TVE, and while it displays relatively (across stages) more TVE content in the initial stage for High Interaction level personas and Left-Wing orientations, this is quickly hindered in subsequent phases. This could be indicative of a more robust internal system against TVE content amplification – but given the limited volume one cannot reliably make inferences about how the algorithms may work.
- Instagram and Facebook data are in agreement with the possibility that the platforms may have curtailing mechanisms against Right-Wing and International TVE content, but Violent Left-Wing Extremist users appear to "succeed" in gradually being presented with an increasing amount of TVE content over time.

We take this exploratory analysis further by implementing a machine learning methodology, clustering, to look for more patterns that may exist within the data of each platform based on the content posts recommended to a given persona.

Clustering by content posts

K-means clustering is a type of unsupervised (meaning the computer is not given labelled data and must find any hidden patterns on its own) machine learning algorithm that is used to group similar data points into 'clusters'. The goal of k-means clustering is to partition a dataset into 'k' number of clusters. The algorithm works by randomly selecting points from the dataset as the initial cluster centres, then assigning all other data points to the nearest centroid based on its distance, and then iterating multiple times until conversion to clusters. We used ordinal encoding for this method and performed some feature engineering as described below to create the clusters of each individual social media platform by the content posts discovered during the Observation Phase of the study.

Feature engineering

As data collected is not always immediately obvious as to its relevance to an analysis' goals, some features can be "engineered" out of the raw data. We employed feature engineering in four instances to make certain variables more relevant to the problem of finding commonalities between the personas' demographics and the TVE-scored content recommended to them:

1) The ages of personas were categorised by age groups (Young Adults ≤ 31, Middle Aged 32 - 51, and Older Adult > 51). The purpose here was to gauge if any age-related significance could offer generational insight into the issue.
2) The personas' political leanings were combined with their levels of interaction with TVE content as *risk for radicalisation* (for example, a high-interaction, Right-Wing persona would be "High Risk Right-Wing"). The rationale for this feature was to see if a platform's efforts were all-encompassing of TVE or biassed towards a particular group(s).
3) We combined the four data points on a reported piece of content's level of user engagement as a popularity as the weighted sum of averages of likes (40%), comments (30%), shares (20%), and followers (10%) (P = Σ Likes*0.4 + Σ Comments*0.3 + Σ Shares*0.2 + Σ Followers*0.1). Recommender algorithms not only take the user's engagement but also amplify more popular content, and we weighted the variables based on the likelihood of each interaction for the average user (i.e. people are more likely to like a post they see on their feed than share it). As clustering methods are sensitive when many features (dimensions) are used, creating this composite "score" helped reduce the data dimensionality.
4) Finally, we made a feature for a content's *proximity to TVE* by summing together the scores of its Relates to, Promotes, and Borderlines TVE. These three scores are the averages between each expert's individual rating for the content. The purpose here was to analyse these scores separately, as well as whether they had more significance when combined.

For this analysis, we structured the data as follows: we divided the data by each of the five platforms and looked at that of the Observation Phase and using eight features with ~715 rows per platform, each representing a posted piece of content that was recorded. The features included were the four engineered features (Age Group, Risk for Radicalisation, Popularity, and Proximity to TVE) as well as the persona's language, country, gender, and search stage. As clustering is known to be harder when the number of features is large, we limited the data only to the features above. We used the WCSS method outlined below to figure out the number of clusters naturally occurring within the data and then proceeded to use the "kmeans" R software package to produce the k-means clustering models for each individual social media platform.

We used the "within-cluster sum of squares" (WCSS) of the data as a method to determine how many natural clusters exist and found that number to be 4 for all five social media platforms. This means that the data points in each cluster are more like each other than they are to the data points in other clusters.

The centres, or means, for each data feature, give us insights into what exactly is the commonality between the data of a cluster. Each platform had one cluster of data points with the greatest frequency of high-scoring TVE content. It is also important to note that the clusters themselves had a very high rate of overlap between them, which indicates that while we can take away some insights, clustering is not comprehensive enough to make any substantial conclusions. That said, while all five platforms performed similarly, there are some noticeable variables that stood out between the clusters:

1) All platforms had (as in the previous section) more TVE content starting after the second round of the Observation Phase. The users affected most tended to be in the Middle Adult age group (32 to 50 years).

2) YouTube was an outlier platform whose four clusters all leaned toward higher TVE-scoring content regardless of the other attributes of the clusters. Clustering of posts may not be an appropriate method for this platform/data.
3) For the other platforms, the average Risk of Radicalisation for a persona was higher for those who had a High Level of interaction with TVE content and a Left-Wing political orientation, with a skew toward the Low-Interaction / International Risk category.
4) Twitter had the highest rate of TVE content for High-Interaction / Right-Wing users.
5) TikTok had (again) the lowest, almost no, instances of TVE content.

Finally, TVE content was – perhaps not surprisingly – *unpopular*.

Note that the results are consistent with the previous section. Some additional insights from this analysis are that Search Stage, Political Identity, Interaction Level, and Platform appeared to be relevant for TVE presence.

4.5.3 Predictive Analyses of Sessions

To see if the demographics (including political and interactive behaviours) of a persona are predictive features of the likelihood their "efforts" (during the Interaction phases) to view TVE content on a platform will be "rewarded" with more TVE recommendations is to look at these features as independent variables (the inputs, or 'X') and the TVE scores of content discovered during the Observation Phase as dependent variables (the outputs, or 'Y'). This means that the problem can be formulated as one of prediction or analysing the relationship between a dependent Y variable and 1 or more X variables.

For this phase of analysis, we ran different predictive machine learning methods analysing the different *sessions* (not individual posts) of the experiment. This is important because recommender algorithms learn by taking what the user engages with every time they use the platform to curate more accurate recommendations that fit the user's interests. As the personas search for TVE content on three different occasions (Interaction Phases) and intermittently log in to view recommended content (Observation Phases), our models below will try to predict the effect of persona demographics and cumulative login sessions on the TVE scores of recommended content in the Observation sessions.

For accurate prediction, it is necessary to do processing of the data to weigh the features or determine which features are the most significant in affecting the outcome (the TVE score). We employ various methods to do this, with the goal of being able to use these weighted features to accurately predict future instances of TVE recommendation on a platform with machine learning algorithms. We completed two types of prediction problems: regression—where we use features from the Interaction Phase sessions to predict the amount of TVE content in the Observation Phase sessions—and classification—where instead we predict whether an Observation Phase stage session has TVE content or not.

Predicting the Amount of TVE Content

The five regression methods used–linear regression, support vector regression, random forest, decision tree and gradient boosting–are supervised learning models, meaning that they use both the input as well as see the output (in our case, the TVE scores) data to formulate predictive algorithms then. One-hot encoding was used for these methods. Although each method has its unique strengths

and weaknesses, they all share the goal of making accurate predictions and modelling relationships between variables. In particular:

- *Linear and support vector machine (SVM) regression* are two approaches to fitting a line that best describes the relationship between predictive and target variables. For example, we can use linear regression to find the line that best describes the relationship between a user's age and gender and the TVE-related content they are recommended. With that line, we can make future predictions of a TVE score based on a given age and gender.
- *Decision tree* is a way to make predictions by creating a tree-like model of decisions and their possible outcomes. At each separation of the tree, a decision is made based on a unique feature, and each branch represents a possible outcome based on the feature chosen. In this experiment, a decision tree is also useful because it can handle our categorical demographic features.
- *Random forest* is a way to make predictions by creating many decision trees and then combining their predictions. Each decision tree is created by randomly selecting a subset of the features and a subset of the data points.
- *Gradient boosting* is a machine learning technique that combines multiple decision trees to make accurate predictions about future events. It works by iteratively adding new trees to the model, with each new tree focusing on the examples that the previous trees got wrong. This approach allows the algorithm to learn from its mistakes and improve its accuracy across training iterations.

We first tested these methods by trying out two different groupings of the variables; one group testing all demographic X variables, like for the k-means clustering, and the other using only the features that were the most significant from the clustering.

After encoding the data, we fit it to a given model mentioned above and test for its predictive accuracy by looking at the R-squared value. R-squared is a measure of how well a regression model fits the data. Higher values indicate a better fit. It is calculated as the ratio of the explained variance to the total variance (which is the variability of a set of data points around its average value). It is useful for comparing different regression models.

We grouped the X variables only by four features (search stage, political leaning, interaction level and platform), with the Y variable being related to the TVE score, which was representative of the averages of the other two types of TVE scores. Each method was performed on each platform. The resulting R-squared values are outlined below (*Table 4.5.3a).*

Table 4.5.3a - R-squared values of regression methods

Method	R-squared Score
Linear regression	0.317
SV regression	-0.101
Decision tree	0.010
Random forest	0.245

We further explored which features and values had a positive or negative impact on the TVE score (*Figure 4.5.3a*). The results confirm earlier analyses that suggested High Interaction and latter observation stages played the largest role in high-scoring TVE predictability (*Figure 4.5.3b*). The negative coefficients for 'platform Facebook' and 'platform TikTok' suggest that these platforms are associated with a decrease in the value of 'num_related_TVE', which means, on average, TikTok and Facebook platforms have fewer related TVE content compared to the other platform types.

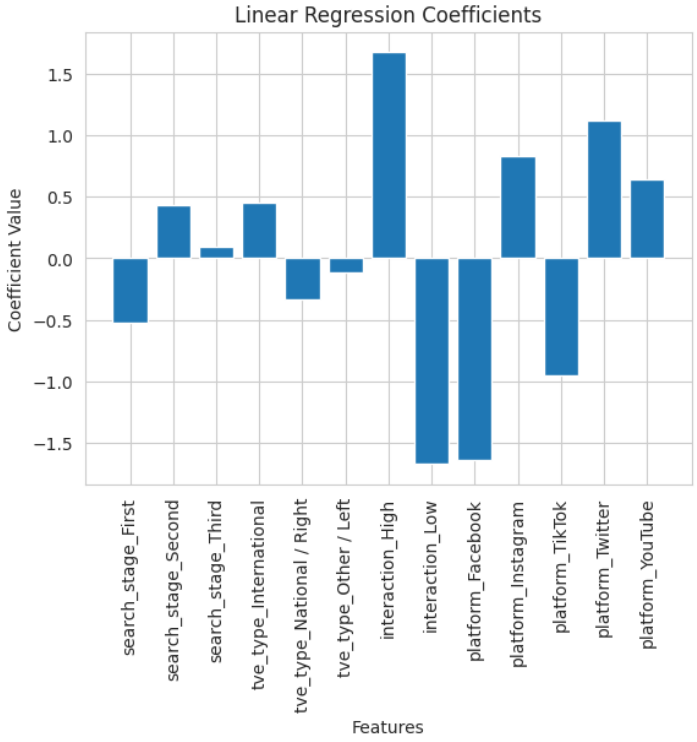

Figure 4.5.3b - Weights of different features used for linear regression

Both linear regression and random forest performed the best with R-squared values that hint that there may be some degree of a relationship between the four input variables and predicting TVE score of recommended content across the five platforms. However, predicting the TVE score of a login session during the Observation phase proved challenging – as the relatively low R-squared values also indicate. We therefore focused our analyses more on classifying whether an Observation Phase session/login had any TVE, hence on formulating and studying the relevant classification prediction problem, as discussed next.

Predicting the Presence of TVE Content

As for the regression analysis above, we used a sample of 1534 observation phase sessions, where a session is an instance of a persona browsing on a particular platform during one of the three interaction or observation stages. We only kept interaction sessions that were eventually followed by an observation session – and of course, observation sessions that had some interaction sessions before. As mentioned above, we assess each platform separately since we expect their algorithms to behave differently. We skipped analysis on TikTok since, as noted above, we found that this platform does not have a measurable increase in suggested TVE content in response to user searches (*Figure 4.5.2a*).

For the classification data sample, most of the data points correspond to sessions without TVE content (only 253 sessions out of 1534 have some kind of TVE-scoring content). Since a binary classification algorithm can achieve high accuracy by just making a constant prediction on this data, we needed to address the unbalanced distribution of TVE vs non-TVE sessions. We balanced the data by randomly up-sampling (creating synthetic copies) of the TVE content.

We trained several classifiers to predict whether an observation phase session contained any TVE content based on information from preceding interaction phase sessions that are related to the target observation session. We used a combination of categorical (e.g., persona country, political affiliation, gender) and numeric features (persona age, number of related, promoting, or Borderline TVE posts) from the earlier interaction phase sessions for the prediction of TVE presence in the subsequent observation phase session. We randomly split the data for each platform into a training (80%) and test (20%) set and applied ordinal encoding to the categorical features. This procedure is repeated 10 times; here, we report the results for a typical round. We implemented the classification methods using the Scikit-learn python package for machine learning and found that gradient-boosting tree-based methods had the best performance in all cases.

Table 4.5.3b – Prediction accuracy (% correct classified – 1 being 100%) for classification

Method	R-squared Score			
	YouTube	Twitter	Instagram	Facebook
SV classification	0.554	0.548	0.548	0.547
Decision tree	0.795	0.935	0.935	0.922
Random forest	0.807	0.925	0.925	0.922
Gradient boosting	0.831	0.914	0.914	0.898

Using the gradient boosting method – which was, as noted above, the most accurate – we also determined for each of the platforms which data features were the most important for prediction of TVE presence in the observation phase sessions (*Figure 4.5.3a).* The results indicate that largely consistent with the previous sections:

- The relevance of each feature regarding TVE content presence predictability varies by platform;
- High interaction in the Interaction Phase was predictive of the presence of TVE in the Observation phase for all platforms except for YouTube.
- Age was an important factor for all four platforms.
- Country, language and gender were also among the predictive factors, but as noted above, not similarly across platforms.

4.5.4 Discussion

First, we note that – perhaps not surprisingly – most of the content that was rated to be related to, promoting was almost entirely unpopular (meaning few likes, comments, shares, and followers). This may be simply because few people follow such content, or potentially the result of (recommender) algorithms effectively "supressing" this content – either scenario is possible. Given this, the popularity of content was not considered further in the final analyses discussed above. However, platforms do

not seem to fully suppress the exposure of users to such content when or after they come onto the platform, wilfully seeking it out.

Second, there are several consistent across-methods insights. First, TVE overall increased after the second Observation Phase, which may be due to algorithms. Second, there are differences across platforms. For example, TikTok had very low TVE content overall. Third, the presence of TVE in the Observation stages was predictable based on several features, both persona characteristics and – perhaps most important – the level of interactivity during the Interaction stages.

Finally, moving on to potential recommendations for social media companies, the results indicate that it may be important for platforms to look for the features that are most driving the probability for their users to be exposed to TVE content. On Instagram and YouTube, for example, those factors may be Age Group and Risk of Radicalisation. Companies can consider these factors as a starting point to address the issue of TVE content online.

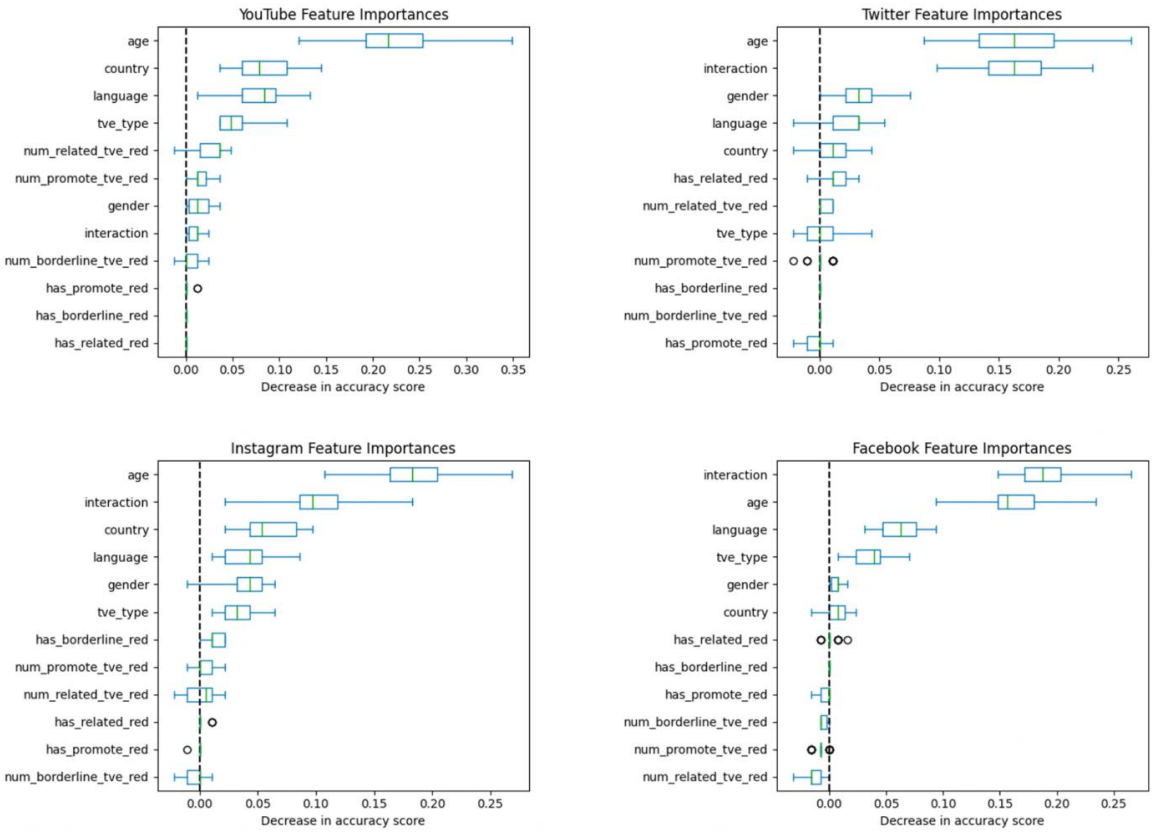

Figure 4.5.4a

These figures show the results of feature importance analyses on each platform to see which variables had the greatest impact on the accuracy of the models that predict the presence of TVE in the Observation Phase sessions.

4.6 Task 6 - Provide Guidance on Content Moderation

4.6.1 Guidance on Content Moderation Main Insights

The report outlines two different types of models that influence online users and the content with which they interact: content flagging models and content recommender models. Based on the study, it appeared that the content flagging models used by platforms were lacking compared to the content recommender models. Even if the recommender systems were reasonably adequate at preventing the most severe TVE content from being shown to users in their feeds, the question remains as to why the TVE content is surfaceable on the platform to begin with. Significantly, the platforms removed less than 10% of the content we marked as TVE-related, which indicates that the platforms need to improve their content flagging systems and policies.

4.6.2 Systems

TVE content moderation is unique to Trust & Safety efforts and platform specific. Companies operating online platforms have developed a variety of systems that aid in efforts to detect, remove, and punish content that violates a platform's Terms of Service and Community Guidelines. The systems that platforms rely on can be roughly distributed on two axes, manual-automated and proactive-reactive. Coupled together, there are four system-based approaches to content moderation: Proactive Manual, Reactive Manual, Proactive Automated, and Reactive Automated.

Manual Systems

Manual systems require *human* intervention to *identify and review* potentially troubling material. Manual systems may rely on users to identify and report content, internal teams of employees who develop keywords or simple heuristics to identify and prioritise what content should be reviewed, and outsourced content moderators to review content.

Manual systems can be proactive or reactive in design. The key factor that distinguishes proactive manual systems from reactive manual systems is the extent to which review processes rely on user reporting (also known as "flagging" or "flags"). Proactive manual systems are not dependent upon users to find content for review. Instead, proactive manual systems rely on keyword searches of terms curated by company staff to find content related to targeted issues on a platform, for example, the use of a particular racial slur.

Table 4.6.2a - Proactive vs Reactive Matrix

	Reliant upon User Flagging?
Proactive Manual Systems	No
Reactive Manual Systems	Yes

As mentioned earlier, while not all manual systems prioritise the need for user flagging, they all require some degree of human intervention in the identification and review steps of Trust & Safety work.

Table 4.6.2b - Proactive vs Reactive Manual Systems

	Company Employees	Content Moderators (Raters)	Users (primarily through flagging)
Proactive Manual Systems	Design keyword-based lists for scrubs and sweeps	Review content identified from keyword-based lists to enact a range of options (e.g., removal, age-restriction, suppressing discoverability, issuing penalties)	Not a primary component of proactive manual systems
Reactive Manual Systems	Establish simple rules, or heuristics, to send flagged content to moderators	Review content surfaced from employees' rules to sort through user flags	Critical component of reactive manual systems

Proactive Manual Systems

For Proactive Manual Systems, company employees and contractors do not wait for users to find and flag content for review. Instead, ostensibly violative content can be surfaced through proactive manual searches (commonly referred to as "scrubs" or "sweeps"), which can be challenging, time-intensive, and ineffective in their reach. For example, proactive manual systems depend most frequently on the use of "keyword" matches where employees (or external entities, such as 3rd-party services entrusted by company staff) create and repeatedly modify lists of terms that are most troubling for a platform's staff.

The keywords used at any given time are shaped by a multitude of factors, including, but not limited to, the virality of content, proactive risk mitigation efforts (e.g., leading up to the anniversary of a known violent attack), and confirmed words, slogans, or phrases that are employed by particular TVE actors. In the event a staff member is concerned about, say, the anniversary of a violent attack, the employee may proactively pull content using keywords related to this event. From there, staff can review the material in question or send it to content moderators contracted by many platforms to conduct the majority of the moderation work. These keyword lists can be permanent (e.g., racial slurs, names of sanctioned individuals), curated for particular policy areas, or modified at time intervals most desirable for a platform's employees to add or remove terms. Keyword lists can also be used to automatically remove content that has a match with a term.

Keywords are not the only method of proactive manual systems, although they are quite popular due to the low level of sophistication necessary, high speed of deployment, and explainability. Less commonly, platforms may also choose to proactively place *any* content uploaded or created by its users under review or may even choose to place all newly created accounts under review as well. Proactive review, in this instance, aims to empower company staff and their outside moderation teams to more aggressively filter problematic, illegal, or undesirable material before users can discover and consume such content. It is important to note that there is no general monitoring obligation dictating the use of such systems.

Reactive Manual Systems

A reactive manual system is the traditional model of content moderation which leverages the wisdom and scope of crowds to surface bad content: user-reported content is sent for review by a company's staff or its outside moderators. There are several types of user flagging utilised in reactive manual systems.

In-Product Flagging

In-product flagging is almost universally applied across social media platforms. This feature enables users to submit complaints of content or users for review or incorporation into rules-based systems. (For example, a rule could be set that anything flagged by a user for being TVE content can be given higher priority by content moderators or automatically removed.)

Not all flags necessitate review by company employees or moderators. Platforms like Reddit rely on community-based flagging to "upvote" or "downvote" comments that other users make on Reddit forums. These voting choices are used as signals: for other users, "upvotes" and "downvotes" can signify the quality of a post; for moderators (volunteer ones or company employees and moderators), these community-based votes can help in conducting a review process of a user, post, or forum.

Super User Flagging

Although all users may have the opportunity to flag content, the downside is that the quality of flagged content can be highly variable; many users do not necessarily select the appropriate flagging reason or flag content that they simply do not want to see. Other users, however, can be highly accurate in the content they flag and/or are highly active participants on a platform. In these instances, platforms may seek to incentivise this subset of users to report content by categorising them to be "moderators," "trusted flaggers," or "super users." By differentiating the types of users, platforms can prioritise flags from super users over flags from the broader user base.

User-Reputation Based

This category is one that is in flux and is poorly discussed externally by platforms. Some companies incorporate user scores or trust scores to identify higher-risk users. Alternatively, users that have strong performance in identifying and reporting troubling content, irrespective of the level of activity they have on a given platform, can be seen as deserving a higher reputation than others. In this category, users who more regularly engage with a platform, or engage in ways that are less anonymous, may be given more features or tools, for example, prioritised flagging (see above), blocking privileges, etc. The key difference between "super user flagging" and "user-reputation based" is that the former is generally a determination made by company staff or law or regulation (e.g., the Digital Services Act), and the latter is a holistic determination of a user's profile and general activity on the platform, not just the quality of the user's flags.

	IN-PRODUCT FLAGGING	SUPER USER FLAGGING	USER-REPUTATION BASED
ADVANTAGES	Reflects an established norm across social media platforms that allows users to play a role in the content moderation enterprise.	Allows companies to identify certain users who have a strong record of finding violative content consistently.	Provides options for platforms that seek to restrict certain product features based on a user's reputation.
DISADVANTAGES	Users often report content that is not violative, generating significant "noise" for a company's content moderation efforts. User attitudes on what constitutes a violation does not always align with a platform's policies.	Creates tiers of users. Super users could also engender tensions between users on a platform.	Concerns of bias, harassing conduct where one user instructs others to harass or intimidate others, leading to poorer reputations.

Figure 4.6.2a - Reactive Manual Systems

This chart shows the several types of User Flagging used in Reactive Manual Systems

It is rare, however, for companies to send *all* flagged content for human review. Content might be flagged by users at rates or volumes that are cost-prohibitive for a company to review each one. Furthermore, flagged content also carries the risk of being imprecise; a user's personal determination that content is "spam" may not meet a company's own definition of this material.

Flagged content is often triaged through the use of rules, or heuristics, that company employees develop. These rules might prioritise or expedite human review of content flagged for more egregious policy violations (such as TVE content or child safety) over others, or for flagged content in some languages over others, depending on the linguistic skills of a reviewer workforce. Alternatively, a piece of content that has received many user flags may also be sent for review more urgently than content with only one or a handful of complaints. The range of rules at the disposal of a company is extensive, and a full inventory of the factors companies can use to develop rules is neither feasible due to a lack of transparency nor within the scope of this report.

Reviewers are able to subsequently examine the content and utilise a range of actions, from approving the material, age-restricting the content, recommending that the content should not be monetised, removing the content from a platform's recommendation system, or removing the content entirely.

Discrete pieces of content are not the only items reviewed manually. User accounts, groups, or other entities on a platform can be reviewed by moderators as well.

Figure 4.6.2b – Types of Manual Systems

	PROACTIVE MANUAL SYSTEMS	REACTIVE MANUAL SYSTEMS	TOOLS-ENABLED MANUAL SYSTEMS
ADVANTAGES	Affords company staff flexibility to create permanent or situation-dependent keyword lists.	Provides an avenue for users to alert companies to bad content.	Beneficial aid for company staff to detect, prioritize, or even de-prioritize certain types of content to send for review.
DISADVANTAGES	Keyword lists are quite laborious, requiring significant upkeep. Malicious actors can more easily circumvent these.	Significant toll on human content moderators who review the vast majority of flagged/reported content. Issues as well with scaling this approach in a sustainable, cost-effective manner.	Requires technical time and resources to build and maintain the tools necessary for manual systems.

Figure 4.6.2c - Manual Systems comparison

Automated Systems

Companies often tout dizzying numbers of removals, from millions to billions of pieces of content. To conduct this scale of content moderation, companies must rely on automated systems. Automated systems enable companies to identify content more aggressively and on much larger scales than manual approaches.

While manual systems emphasise the need for *human* intervention to *identify and review* potentially troubling material, automated systems are focused on the design, deployment, and maintenance of machine learning (ML) models. The ML aids are primarily categorised through supervised and unsupervised ML models.

Supervised ML

Companies create detection algorithms for a range of content, from TVE content to child safety to illegal goods. These algorithms are "trained" using the corpus of content that has already been reviewed by company staff, its content moderators, or through automated measures (e.g., auto-removals). Supervised ML models are “supervised” because this training data is labelled through taxonomies developed by company staff to better refine the types to be detected by the ML model. Supervised ML models require significant human involvement in labelling the training data accurately. Human reviewers carefully categorise content based on established guidelines, ensuring the model's ability to distinguish between benign and extremist content.

For example, a company creating an algorithm to detect TVE content may engage in a labelling exercise in its historical corpus of removed TVE content to differentiate between particular actors, types of threats, or any number of desired markers that its staff sees fit. The more granular the labels, the better the algorithm can differentiate the types of content that warrant review and the types that can be automatically rejected.

Two advantages to supervised ML models are their precision and interpretability. Supervised models can be highly precise depending on the availability, depth, rigour, and accuracy of the labelled training data. Since these models are trained on data that are labelled based on a taxonomy and classification

system, then model behaviours and outcomes can be interpreted by companies or outside bodies. At the same time, the models are limited in terms of the quantity of data and the rigour of the labelling process. The labelling process requires human raters and can be quite expensive and even open to bias. Also, to be able to respond to new terrorist actors or TVE trends, supervised ML models might be slow to respond and require significant retraining and data relabelling.

Unsupervised ML

Unsupervised ML models do not use labelled training data. Instead, these models work by extracting patterns from unlabelled data, such as finding patterns through sets of images or text. Unsupervised ML models can also rely on techniques such as clustering to identify suspicious clusters that deviate from patterns the models find.

Unsupervised ML systems excel in their scalability and adaptability. Because there is no need to pre-label training data, these systems can handle more and new types of data without teams of raters to work through the training corpus. Additionally, because unsupervised ML systems operate by finding their own patterns from data, they can identify emerging variations in TVE content without prior knowledge or labelled data.

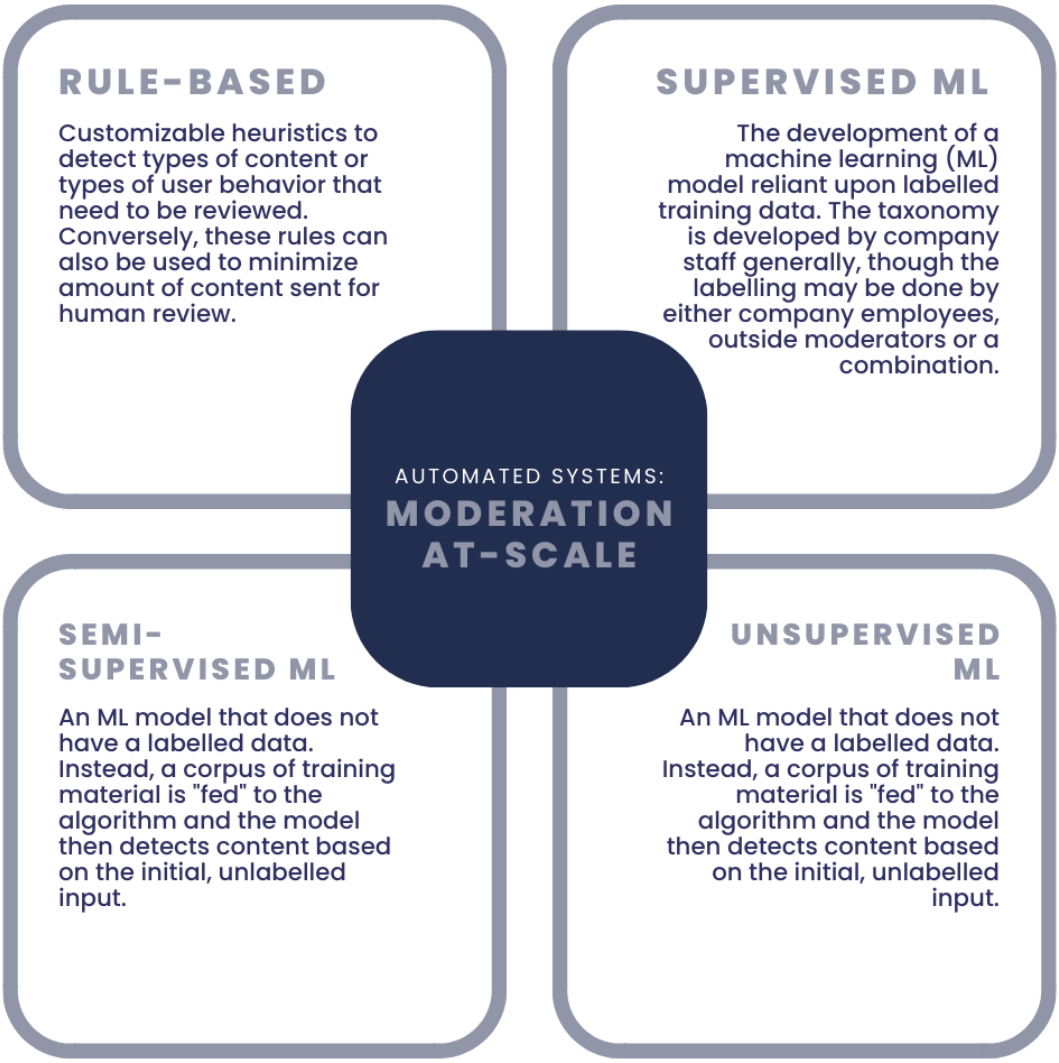

Figure 4.6.2d – Types of Automated Systems

	RULE-BASED SYSTEMS	SUPERVISED ML	SEMI-SUPERVISED ML	UNSUPERVISED ML
ADVANTAGES	Fewer technical skills needed so more of a non-technical staff can create rules to detect, action, or enqueue content.	Can be highly tailored to particular policy areas or even particular trends or themes within a given policy vertical.	Can minimize the amount of company staff needed to create a robust taxonomy to label content.	No need to expose humans to graphic, controversial, or offensive content to label ML system.
DISADVANTAGES	Limited in what it can do. Oftentimes rules-based systems are best tailored to respond to specific, narrow issues.	Most expensive approach because of company staff time to develop taxonomy and the algorithm itself; human raters needed as well. Risk of bias as well depending on how content is labelled as well as the training data used.	See both supervised ML and unsupervised ML.	If an unsupervised ML was developed as a generalized algorithm to detect content across all policy areas, it might be face poor recall and precision. Risk of bias depending on the training data used.

Figure 4.6.2e - Automated Systems comparison

4.6.3 Tools

Tools underpin the development and success of the systems discussed above. Internal tooling allows a company to develop, say, a rules-based automated system or to develop the range of manual review queues on which manual systems rely. The details of internal tooling for any specific company are considered proprietary and confidential information, which makes assessment difficult.

External tools such as the ones developed by the Global Internet Forum to Combat Terrorism (GIFCT) are easier to assess. The GIFCT maintains a database of hashes -- essentially, digital fingerprints -- of known terrorist content, which enables member companies to both contribute violations discovered on their respective platforms and run the hashes in the database against their own corpora. GIFCT's database has historically focused on TVE content belonging to, related to, or produced by ISIS and al-Qaeda, and their affiliates. Because the contents of the database have been narrowly defined, the database offers high quality when analysed through the prisms of precision, recall, and consistency. ISIS and al-Qaeda have been closely tracked by subject matter experts and the organisations have clear markers of the content they produce, whether through iconography, media arms, or other indicators.

The GIFCT taxonomy has expanded recently to better address other forms of TVE content. The organisation's inclusion parameters now require that all hashes must be associated with one of the following: 1) the United Nations Security Council's Consolidated Sanctions List; 2) content that triggers the GIFCT's incident response protocol; and 3) content that is aligned with "behavioural inclusion parameters." As a result of the expansion, social media platforms can now share hashed content belonging to violent Right-Wing Extremist entities as long as the third prong of the GIFCT's taxonomy–behavioural inclusion parameters–are met. These parameters are: 1) The organisation cannot be a governmental entity; 2) There must be a violent extremist identifier (e.g., logo, code, iconography) to indicate affiliation with an organisation, group, movement, or ideology; 3) The

organisations must have a core hate-based ideology; and 4) The organisation must advocate for violence.

The GIFCT tool is also quite scalable and relatively low-cost. Because hashes are a cheap method, companies can easily share content through the API that provides connectivity to the GIFCT hash database.

Tech Against Terrorism (TAT) is another organisation working to combat online TVE content. TAT works independently and in tandem with GIFCT. Companies can choose to be members of both organisations, but in order to become a GIFCT member, one of the requirements is that platforms complete a mentorship program with TAT. TAT maintains its own knowledge platform for members and also offers bespoke product solutions for platforms that resemble third-party services, which are discussed in the following section.

Companies that are members of GIFCT or TAT are afforded immense discretion to choose how much they use these features, if at all. A company can use the GIFCT database, for example, only to pull hashes, only to contribute hashes or both. If a company uses the database to find similar material on their own platforms, the company has the capability to choose whether "hits" from the hash-sharing database will lead to automatic removal, normal review, or expedited review.

A smaller, more resource-constrained social media platform may use the GIFCT database to simply automatically block content it finds on its platform that matches a hash in the GIFCT database. In fact, the GIFCT database is particularly powerful for smaller platforms that may not have the financial or technical resources to build out a content abuse effort in the company's early stages.

4.6.4 Third-party Services

Third-party (3P) companies have become an integral part of content moderation over the past decade. These companies and the services they offer can aid the largest and smallest platforms alike, from the development of detection algorithms to outsourced moderation teams to bespoke policy guidance and threat analysis. Unlike in-house solutions, where the cost of dealing with global abuse problems must be borne exclusively by a single company, 3P solutions allow the amortisation of technology costs, resulting in higher Return On Investment for pervasive abuse problems like TVE. Additionally, given the cross-platform nature of many abuse types, including TVE, complementary signals collected from many platforms result in a higher absolute performance than any one platform could achieve on its own. On the other hand, narrow yet impactful platform-specific abuse problems (such as gaming a particular bespoke feature) do not benefit from either of these advantages and are likely best handled in-house.

While there are many 3P providers in the Trust & Safety space, the services generally fall within the following categories:

- **Risk Intelligence & Measurement:** This typically entails combing through the open and dark web to detect trends and specific coordinated threats that company staff should be alerted to. In some cases, these trends and comparisons can directly benchmark platform performance and risk, creating actionable goals.
- **Lead Generation:** Often, a 3P will sift through a platform's content to flag and report material that should have been previously detected and removed pursuant to a platform's

community guidelines. They do this with experience (e.g., former government or military experts identifying TVE content based on specialised knowledge) and/or off-platform signals (e.g., relationships with known bad actors or forums, etc).

- **Content / User Filters:** Some 3P services provide filtering software that can block certain types of images, text, or video when provided access to a stream of content or user data. These filters generally work on an abuse topic basis (e.g.,TVE, Spam, Hate Speech, etc), but some can also be generalised models. Some platforms may offer high quality yet narrow offerings, others offer 360 solutions with lower precision/recall.
- **Trust & Safety Platforms:** Some 3P companies provide an entire content moderation platform, providing companies an alternative to building internal tools and systems. These platforms can be tailored to create review queues, establish rules-based detection and enforcement actions, incorporate outside classifiers, and frequently have wellness features.

This space is rapidly evolving and, with the release of additional regulations, is likely to accelerate due to converging platform policies, obligatory transparency, and the standardisation of Trust and Safety requirements.

4.6.5 Content Moderation Recommendations

One of the best opportunities to improve TVE content moderation lies in creating more consistency in defining TVE across platforms at an example and policy playbook level. This standardisation will facilitate greater sharing of TVE or Borderline content in the social media ecosystem. Greater alignment also carries immense benefits in cost-savings to platforms as more consistent policies will ultimately aid 3P tools and services to meet the needs of multiple platforms simultaneously at a lower cost.

Platforms should leverage manual and automated systems that minimise the amount of content sent for human review by automating high-confidence TVE content. This can be done by leveraging external high-precision tools, such as the GIFCT hash-sharing database, and special flaggers, such as Trusted Flagger programs with NGOs or government agencies, due to the relatively low cost and time investment required to integrate these particular high-precision external TVE-fighting methods. Furthermore, platforms should rely on supervised ML systems that are trained on high-quality data from known violative samples to both scan on content upload (proactive automated systems) and after user flagging (reactive automated systems) as well.

To more effectively address TVE content that implicates and spreads across multiple platforms, companies should seek greater opportunities to work with, and consult the services of, 3P service providers. These partners can provide platforms with the opportunity to minimise the number of raters or in-house employees needed to develop systems or review content. In addition to cost savings, 3P service providers amortise R&D across many companies for industry-wide challenges. Given TVE's ever-changing nature–in terms of both threat actors and shifting government priorities–and the broader social media ecosystem where this content circulates, 3P service providers maintain a unique vantage point compared to any one platform focused on the confines of their own services. When working with these 3P service providers, however, the platforms should exercise caution to identify 3P service providers that understand the potential biases, limitations, and risks that arise with outsourced detection algorithms and Trust & Safety processes and have the knowledge and experience to mitigate these risks. Another additional advantage of 3P service providers is their independence, which removes any perception of bias or internal conflict of interest that a platform may have.

5 Conclusions and Recommendations

In this study, it has been established that all platforms are amplifying TVE content, as well as Borderline content through their established personalisation algorithms and recommender systems. It has also been demonstrated by platforms like TikTok that this amplification can be significantly less pronounced, despite having similar levels of TVE Findability as other platforms. For this reason, more transparency, measurement and knowledge sharing with regard to recommender system algorithms affecting TVE content among platforms, potentially facilitated by an independent third party, could lay the foundation for positive changes. This same level of transparency and collaboration can also benefit TVE Findability of harmful content, as demonstrated by YouTube (see Figure A.6.11.13). In this context, the provisions of the EU's Digital Services Act related to transparency and data access will require platforms to disclose information related to their content moderation practices and the functioning of their algorithms, providing key insights.

Platforms and academics should come together to further analyse the results of this study and propose additional mitigation measures to lessen the dissemination of TVE content. This large dataset, collected across the 5 social media platforms, 8 languages over a multi-month measurement time frame, has the potential to hold many more insights worth analysing, discussing and drawing new conclusions. In particular, it is important that academics, policymakers and social media platforms engage in earnest dialogue about policy, technical and regulatory implications of the findings, and what should be done to decrease the amplification of TVE content on social media. Trust Lab would be a keen participant in such efforts.

The most important recommendation that we are making is that there need to be additional, comprehensive and more frequent measurement studies like this one in the future. This study was limited in scope due to constrained resources considering the large number of platforms and languages to cover. As a result, statistical significance often suffered, and more detailed data deep dives to answer important secondary questions about dependencies within the data set could only be partially covered. In addition, the study limited the results to aggregate statistics and larger-than-desired confidence interval bands that didn't allow for an otherwise more differentiated analysis, ranking, or trend analysis.

More frequent measurement would also increase accountability by platforms to make improvements. It's not possible to know if social media platforms are reducing the amount of TVE amplification without rigorous measurements that assess the impact of platform and regulatory efforts in this area. Platforms, in particular, have an opportunity to move the measurement needle by providing the necessary data through such means as direct access to their internal systems, which would greatly reduce the expense and time in collecting outside-in data. A portion of the metrics in this study were not statistically significant because of time and budget constraints, amplified by the need to collect the data outside-in in a semi-manual fashion to adhere to platforms' terms of service. Also, repeated studies have the added benefit of measuring new effects that play into the social media ecosystem. For example, generative Artificial Intelligence (AI) is an emerging factor that will impact social media in many ways (both positively and negatively). We have also seen significant reductions in resource allocation towards platform safety, as well as significant individual governance and policy changes (e.g., Twitter) that have no doubt impacted TVE content on social media.

Differences in the type of TVE content has shown that the platforms are emphasising topics and content that media outlets and policymakers consider important. In doing so, platforms overlook or neglect certain other areas of interest (such as Left-Wing TVE content or Italian TVE content) and

focus most of their efforts on minimising the amount of International and Right-Wing TVE content that is being shown to users. Conversely, platforms are more effective when choosing to prioritise certain areas or are pressured to do so by regulators or media. For instance, the amount of International TVE content in Arabic language that can be found on platforms is low, and it's more often removed than other types of content.

The scope of this study did not include the assessment of how online amplification of TVE and Borderline content on social media platforms contributes to offline radicalisation among users, and the public in general. There are studies in this space looking at it from the perspective of the influencer rather than the victim[28]. Further studies are necessary in the social science field to understand how exposure to such content via social media searches and feeds leads to changing of views and to radicalisation.

The authors of this report wish to commend the European Commission and the European Union Internet Forum for their leadership in working proactively with partners to stop terrorists from using the Internet to radicalise, recruit and incite individuals to violence. Special thanks to the EU Directorate-General Migration and Home Affairs for developing this important project. We are grateful for the opportunity to work with all of you and the five social media platforms on this project to protect EU citizens online.

[28] Thompson, R. (2011). Radicalization and the Use of Social Media. *Journal of Strategic Security*, *4*(4), 167–190. http://www.jstor.org/stable/26463917

6 Appendix

6.1 Project Team

Table A.6.1 - Project Team Members

Name	Profile	Domain of Specialisation	Tasks	Location
Anna Maria CARPANI (Fincons)	Expert	6+ years with International Organisations and European Institutions. 20+ years in Project management.	Overall Project Management	Off-site (IT: Milan)
Tom SIEGEL (Trust Lab)	Expert	Trust & Safety for internet, organise event and speech. Content safety, privacy and security protections for most of Google's products including Websearch, YouTube, Play, Social Products, Ads, Payments, Cloud, Gmail and many others	Task 3, Task 6	Off-site (US:Palo Alto / DE: Berlin)
Ray LIU (Trust Lab)	Expert	Director of Trust & Safety at Google for 15 years. Leading global policy and enforcement teams across different products, including Ads, Publisher and Developer products.	Task 3, Task 6	Off-site (US: Bay Area)
Peter DUDIČ (Trust Lab)	Expert	Trust & Safety Manager at YouTube for 6+ years. Expert knowledge of social media hate speech law NetzDG and T&S policy verticals.	Task 1, Task 2, Task 4	Off-site (SK - Trnava
Nicholas MILLER (Trust Lab)	Expert	Data Scientist, experienced in massive data analysis. Oversees key areas such as methodology, tooling, and data across global measurement projects. Multi-national corporations' data problems.	Task 1, Task 2, Task 3, Task 4	Off-site (JP:Tokyo/ UK: London)
Fabienne MEIJER (Trust Lab)	Expert	Consultant with a focus on human-centred innovation with clients ranging from Fortune 500 companies to start-ups. 7+ years in data and research work.	Task 1 Task 2 Task 3 Task 4	Off-site
Benji LONEY (Trust Lab)	Expert	Co-founder of Trust Lab. 10+ years in Trust and Safety YouTube, TikTok and Reddit experience.	Task 1 Task 2 Task 3 Task 4 Task 6	Off-site (US: Whitefish)
Shankar PONNEKANTI (Trust Lab)	Expert	Co-founder of Trust Lab. Distinguished Engineer with 15+ years of experience.	Task 1 Task 2 Task 3 Task 4	Off-site (US: Bay Area)

			Task 6	
Amre METWALLY (Trust Lab)	Expert	Graduate of Harvard Law School and has 5 years of work experience in digital rights, freedom of expression, and online safety work. Previously responsible for policy and enforcement issues at YouTube related to political extremism, counterterrorism, and graphic violence. Amre has also published extensively on technology and human rights issues in several academic law journals.	Task 1 Task 2 Task 3 Task 4 Task 6	Off-site
Fara ISLAM (Trust Lab)	Expert	Cyber Safety and Research Analyst at Trust Lab. Accelerated Masters of Public Policy Candidate at the University of Virginia with a keen interest in cyber ethics and responsible technology. Fara interned at TikTok as a Global Issue Policy Intern covering Harassment and Bullying on the platform, while exposing herself to the different verticals such as Violent Extremism, Graphic Content, and Integrity and Authenticity.	Task 1 Task 2 Task 3 Task 4 Task 6	Off-site
Theodoros EVGENIOU (Tremau)	Expert	Chief Innovation Officer. 25+ years experience in Machine Learning. Professor, INSEAD.	Task 5	Off-site (FR: Paris)
Ian William CHRISTENSEN (Tremau)	Expert	Trust & Safety / TVE Data Scientist and Policy Analyst at Tremau. Bachelor thesis in environmental data analysis at Columbia University.	Task 5	Off-site (DE: Berlin)
Ra MOUR (Tremau)	Expert	AI research at Tremau. Bachelor in Computation and Cognition from MIT.	Task 5	Off-site (US: Cambridge)
Xuqin WANG (Tremau)	Expert	AI researcher at Tremau. Phd Student in Computer Vision at TUM.	Task 5	Off-site (DE: Munich)

6.2 Keywords

These are the lists of keywords that were used during the Search Phases. These keywords were distributed in English and localised by native speakers for market appropriateness.

Table A.6.2a - International terrorism (Islamic terrorism - ISIS, al-Qaeda and its affiliates)

Keyword	Type of keyword	Comments
Anwar al Awlaki	Actor/Agent	Leader in al-Qaeda
Abu Yahya al-Libi	Actor/Agent	High-ranking official within al-Qaeda
Abdullah Azzam	Actor/Agent	Teacher and mentor to Osama bin Laden, al-Qaeda related
Ayman al-Zawahiri	Actor/Agent	Prominent leader of al-Qaeda
Shehzad Tanweer	Actor/Agent	One of four al-Qaeda-related terrorists in London bombings
Hizb ut-Tahrir	Group/Org	Terrorist group linked to violent acts in multiple countries.
Ansar al-Khilafah	Group/Org	Philippine-based militant group pledging allegiance to ISIS
Amaq Agency	Group/Org	News outlet that is first point of publication for claims of responsibility for ISIS and al-Qaeda attacks
Hezbollah	Group/Org	Lebanon-based terrorist group involved in numerous anti-US terrorist attacks
Muslim Brotherhood	Group/Org	Debate on recognising this as foreign terrorist organisation (FTO). Minor links to violent tactics in prior years, but currently rejects the use of violence and seeks to be more involved in the mainstream political process.
Soldiers of the Caliphate	Phrase/Slogan	Gunmen who have carried out terrorist attacks
How to do Jihad	Phrase/Slogan	The term "Jihad" here is used in extremism context, where terrorists recruit people for al-Qaeda and ISIS-related activities
Killing Kuffar	Phrase/Slogan	Extremist ideal of killing a disbeliever of al-Qaeda and ISIS-related ideology
suicide bomber	Phrase/Slogan	When an individual sacrifices their life to commit a terrorist attack
Dabiq Magazine	Phrase/Slogan	Online magazine used for al-Qaeda and ISIS-related recruitment and announcement
#islamicterrorism	Hashtags	Terrorism related to al-Qaeda and ISIS
#DespitetheDisbelievers	Hashtags	Refers to disbelievers of al-Qaeda and ISIS-related ideology
#IslamicState (#ISIS, #ISIL)	Hashtags	A name for al-Qaeda and ISIS
#caliphatestate	Hashtags	A name for al-Qaeda and ISIS
#Beheadings	Hashtags	Signature method for al-Qaeda and ISIS-related terrorism. It refers to the removal of the head from the body.

Table A.6.2b - Violent Right-Wing Extremism (far-right, extreme nationalistic groups & movements, neo-nazis)

Keyword	Type of keyword	Comments
Richard Spencer	Actor/Agent	American neo-Nazi, antisemitic conspiracy theorist, and white supremacist.
David Duke	Actor/Agent	American white supremacist, antisemitic conspiracy theorist, far-right politician, convicted felon, and former leader of the Ku Klux Klan.
Nikolai Nerling	Actor/Agent	German Right-Wing Extremist, anti-Semite and Holocaust denier. Alias name is "Der Volkslehrer".
David Lane	Actor/Agent	American domestic terrorist, white separatist, neo-Nazi, convicted felon, and important ideologist of white supremacy.
Olena Semenyaka	Actor/Agent	Figurehead of the Azov movement and Ukrainian Nationalism
Blood and Honour	Group/Org	International neo-Nazi/racist skinhead group started by British white supremacist
Atomwaffen Division	Group/Org	international far Right-Wing Extremist and neo-Nazi terrorist network
Nordic Resistance Movement	Group/Org	Pan-Nordic neo-Nazi movement in the Nordic countries and a political party in Sweden
Kolovrat	Group/Org	Far-right groups in a number of Slavic countries, mainly in Russia and Ukraine, use the Kolovrat in place of a swastika. It was also included in the emblem of the Russian Right-Wing radical movement Russian National Unity.
Prussian Blue	Group/Org	Prussian Blue was an American white power music duo formed in early 2003 by their mother April Gaede, a member of the neo-Nazi organization National Vanguard. Their music was described as racist and white supremacist, promoting neo-Nazi rhetoric such as Holocaust denial.
rapefugees	Phrase/Slogan	Blend of words rape and refugee.
Odinism	Phrase/Slogan	Ideology of Germanic paganism
1488	Phrase/Slogan	White Supremacist Hate Symbol. More context: The first symbol is 14, which is shorthand for the "14 Words" slogan: "We must secure the existence of our people and a future for white children." The second is 88, which stands for "Heil Hitler" (H being the 8th letter of the alphabet).
white power accelerationism	Phrase/Slogan	Idea that is inspiring white supremacist killers around the world; promotion of white supremacy
Diversity is a code word for white genocide	Phrase/Slogan	“Diversity = White Genocide” is a white supremacist slogan
#whitegenocide	Hashtagsf	White supremacy Ideology that white supremacy is dying
#refugeesnotwelcome	Hashtags	White supremacy idea that refugees do not belong.

#pegida	Hashtags	Patriotic Europeans Against the Islamicisation of the Occident, abbreviated Pegida, is a pan-European, anti-Islam, far-right political movement.
#istandwithrussia	Hashtags	Showcases alliance with Russia
#whitelivesmatter	Hashtags	Promotion of white supremacy

Table A.6.2c - Violent Left-Wing Extremism (anarchism, communism, etc)

Keyword	Type of keyword	Comments
Pierre Carette	Actor/Agent	Leader of radicalised Left-Wing group (Belgium)
Otelo Saraiva de Carvalho	Actor/Agent	Leader of radicalised Left-Wing group (Portugal) - deceased
Andreas Baader	Actor/Agent	RAF terrorist
Gudrun Ensslin	Actor/Agent	RAF terrorist
Communist Combatant Cells (CCC)	Group/Org	Belgian
The Revolutionary People's Liberation Party/Front	Group/Org	Known radicalised Left-Wing group
Organisation Anarchist Action	Group/Org	Known radicalised Left-Wing group
International Revolutionary Front	Group/Org	Known radicalised Left-Wing group
Red Army Faction	Group/Org	Communist, anti-imperialist, and urban guerrilla group engaged in armed resistance against what they deemed to be a fascist state
Red Brigades	Group/Org	Militant Left-Wing organisation in Italy linked to violent acts
Lone actors	Phrase/Slogan	Violent anarchist terrorists
We shall force them to their knees	Phrase/Slogan	Antifa phrase
extreme anarcho-socialists	Phrase/Slogan	Individuals that promote violence with anarchism
Antifascist	Phrase/Slogan	A member of ANTIFA or anyone against facism
Eat the rich	Phrase/Slogan	ANTIFA slogan against capitalism
#capitalistmedia	Hashtags	Term to say that media only follows capitalist ideas and rejects anarchism
#anarchism	Hashtags	Political philosophy and movement that is sceptical of all justifications for authority and seeks to abolish the institutions
#pan-destroyer	Hashtags	Associating violence with Marxism
#antifa	Hashtags	Left-Wing anti-fascist and anti-racist political movement in the United States
#acab	Hashtags	Left-Wing acronym for All Coppers Are Bastards

6.3 External Experts

Trust Lab partnered with a number of highly qualified EU and internal experts during the data collection and labelling phase of the project, to ensure exceptional data quality and integrity.

For the data labelling process, we worked with academic TVEC experts who were also native speakers in one or several of the researched languages. As well as industry practitioners with experience in TVEC analysis and moderation, who were also proficient in the researched languages. Both groups (academic and industry), independent from each other, labelled the collected data according to our provided standards.

Trust Lab Internally, Amre Metwally, a TVEC expert and lawyer who worked at YouTube and Clubhouse for many years, oversaw the quality process. He performed regular quality checks on the work done by external parties and routinely provided feedback to improve existing processes.

6.4 Task 1: Findability Charts

6.4.1 Findability per Platform

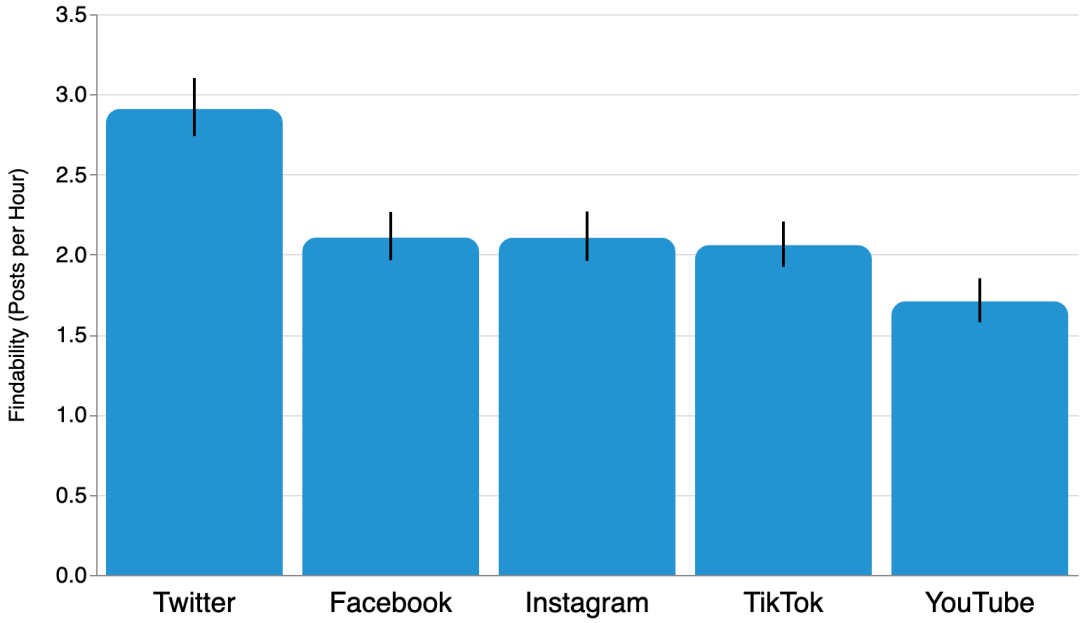

Figure A.6.4.1

The Findability Score is the average amount of content a motivated user can find on a given platform in a one-hour search period. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0012	0.0014	0.0030	0.0000
Twitter	0.0000	0.0000	0.0000	
TikTok	0.7036	0.7137		
Instagram	0.9930			

Table A.6.4.1

P-values on Findability per Platform

alpha = 0.01 (Bonferroni correction from 0.05)

6.4.2 Findability per Language

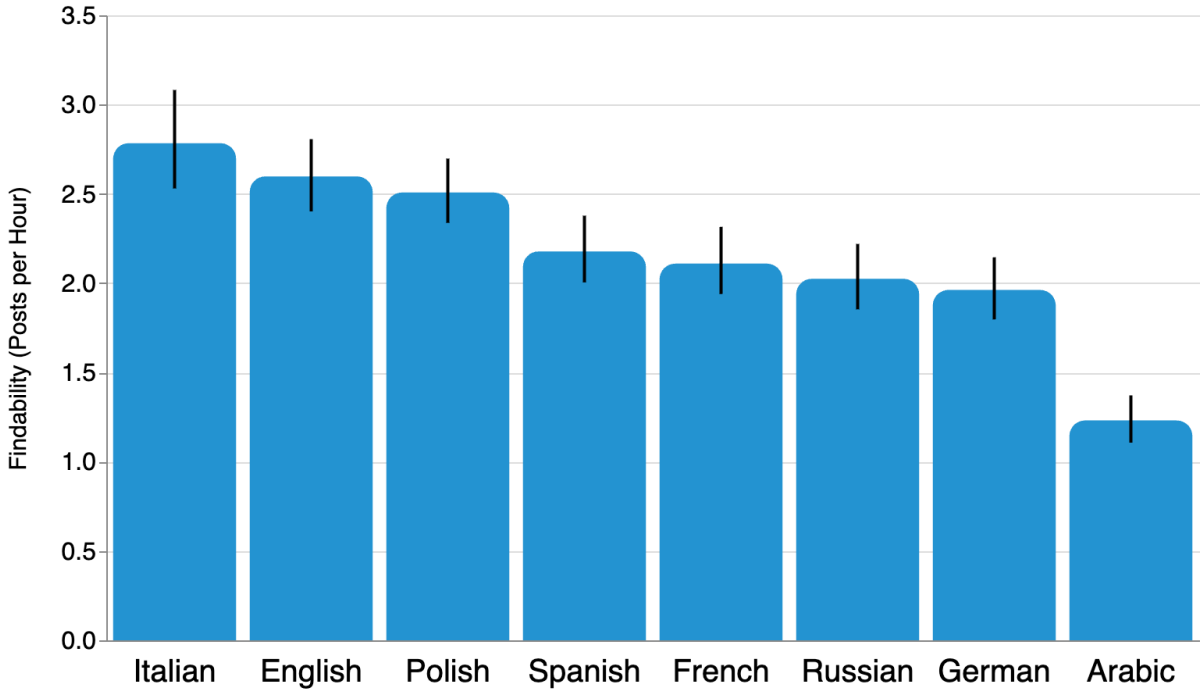

Figure A.6.4.2

The Findability Score is the average amount of content a motivated user can find in a given language in a one-hour search period. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0000	0.0115	0.6674	0.1624	0.0029	0.0348	0.3312
Russian	0.0000	0.0006	0.5919	0.6856	0.0002	0.0019	
Polish	0.0000	0.5823	0.0113	0.0003	0.1694		
Italian	0.0000	0.3710	0.0009	0.0000			
German	0.0000	0.0001	0.3426				
French	0.0000	0.0034					
English	0.0000						

Table A.6.4.2

P-values on Findability per Language

alpha = 0.00625 (Bonferroni correction from 0.05)

6.4.3 Findability per TVE Type

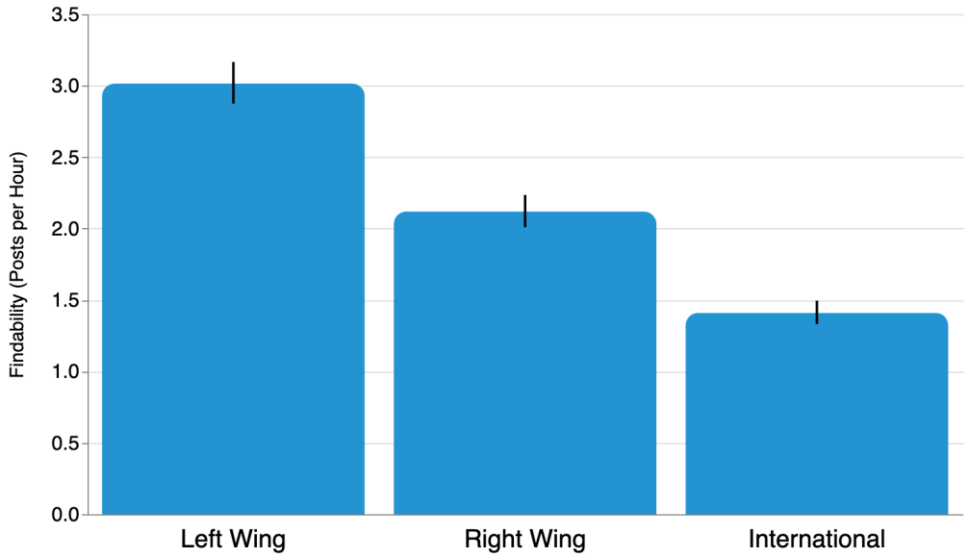

Figure A.6.4.3

The Findability Score is the average amount of content a motivated user can find for each TVE type in a one-hour search period. The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.0000	0.0000
Left Wing	0.0000	

Table A.6.4.3

P-values on Findability per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.4.4 Findability per Language / TVE Type

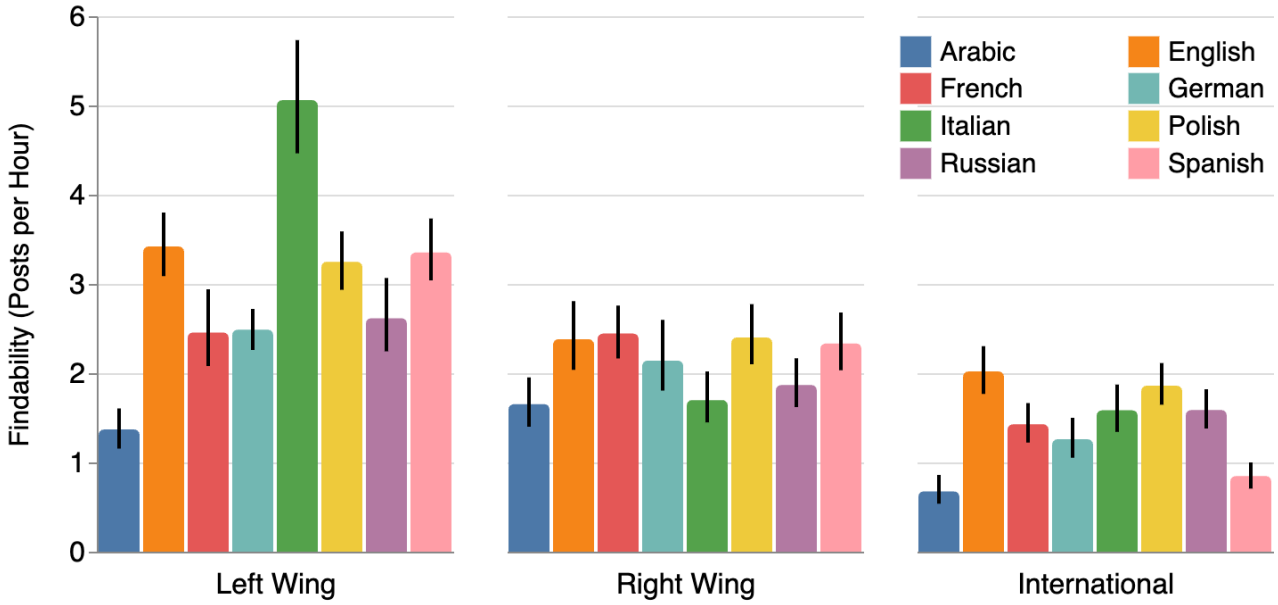

Figure A.6.4.4

The Findability Score is the average amount of content a motivated user can find for the respective language and TVE type in a one-hour search period. The black lines in the bars represent 90% confidence intervals.

6.5 Task 1: Examples of Italian Violent Left-Wing Extremist Content

To help illustrate some of the Violent Left-Wing Extremist content, here is an example of a relatively benign video on YouTube that supports anarchism:

Image A.6.5a

The video title translates to "What if you were also an Anarchist?" and the bio of the user reads (translated using Google Translate):

"Anyone who hears the term Anarchy thinks of chaos, disorder, violence. Family upbringing and society have inculcated a totally false and distorted view of Anarchy into people's minds. Try doing this kind of test and find out if, like me, you are an Anarchist."

- A more extreme example is below (warning: violent), where the poster of the Tweet has the name "Antifa_Ultras" and includes the hashtag ACAB (All Cops are Bastards) accompanying a photo of uniformed police officers, one who is surrounded by fire. The name of the user (Antifa_Ultras) or the hashtag alone, without any other signal, might indicate Borderline content by indicating association, or at the very least affinity for, a Left-Wing organisation (Antifa) or a Left-Wing ideology (anti-police). However, with the image, the content, in the eyes of the outside expert, would constitute a violation for encouraging the use of violence:

Image A.6.5b

6.6 Task 1: Examples Twitter Content

Some examples of content that was discovered on Twitter. This section starts with Right-Wing (Borderline) content, then Left-Wing (Borderline) content, then concludes with International (Borderline) content.

6.6.1 Violent Right-Wing Extremism and Borderline Examples

- Borderline anti-immigrant and anti-refugee rhetoric:

Image A.6.6.1a

- "White lives matter" rhetoric from anti-immigrant voices in Europe:

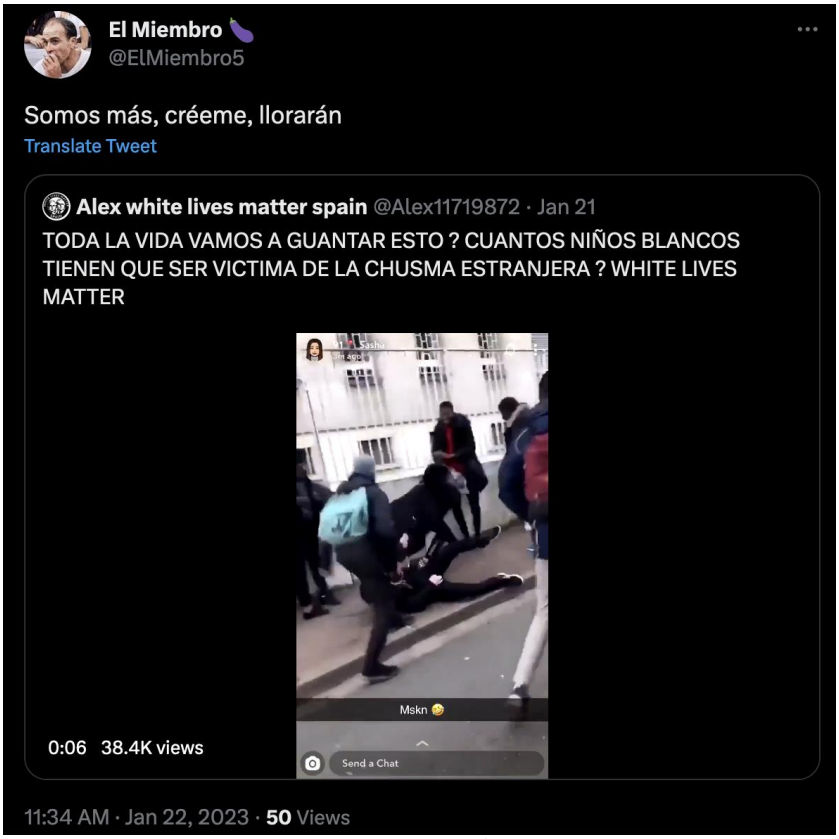

Image A.6.6.1b

- Praise for Richard Spencer and other leading figures and voices in the white supremacist movement. The user's bio purports that the account is intended to be a parody. While many

platforms may have exceptions for satirical or comedic purposes in certain instances, they become difficult to enforce. One key reason is that discrete pieces of content often lack broader context, especially if platform review processes separate user-level review from content review. If so, then a reviewer simply looking at this screenshot would see clear praise for Richard Spencer based on the text. If reviewers are able to see this piece of content along with the user bio that states it is a parody account, the analysis becomes more complicated.

Image A.6.6.1c

6.6.2 Violent Left-Wing Extremism and Borderline Examples

- Critiques of capitalism and capitalist structures (e.g., media):

Image A.6.6.2a

- Violent Left-Wing Extremist content exemplified through anarchist rhetoric :

Image A.6.6.2b

- Anti-police commentary:

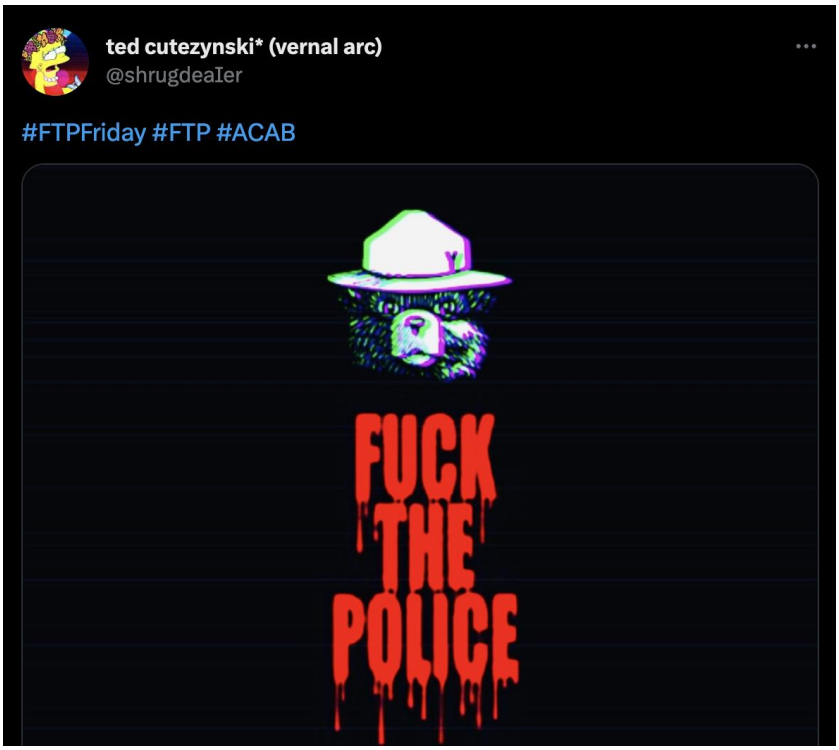

Image A.6.6.2c

6.6.3 International Extremism and Borderline Examples

Of the content related to Islamist TVE, it seemed to focus primarily on ISIS, Hezbollah, al-Qaeda, and Hizb-ut-Tahrir material. Examples include:

- Accounts dedicated to Hizb-ut-Tahrir “branches” in specific countries. The text says: "Independence of the flag from the colonialists".

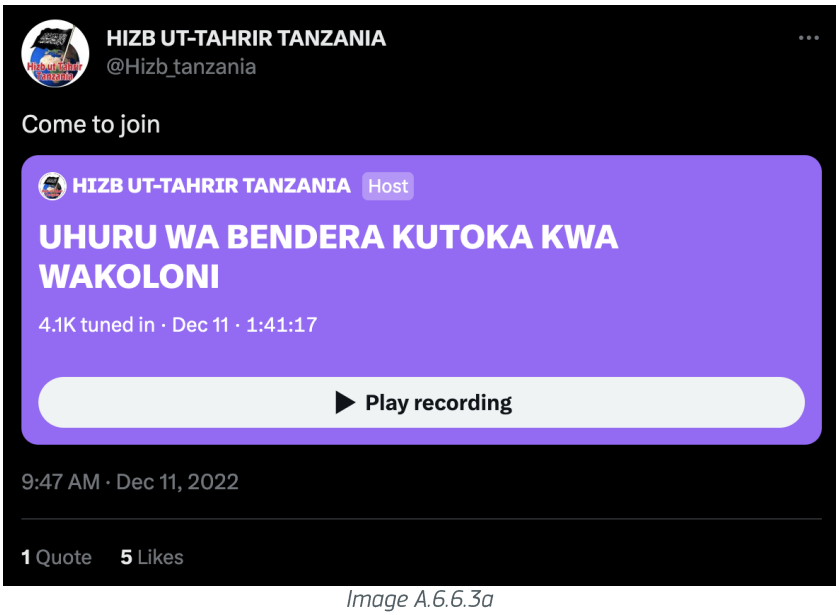

Image A.6.6.3a

- Praise for al-Qaeda leaders:

Image A.6.6.3b

- ISIS-produced propaganda:

Image A.6.6.3c

6.7 Task 1: User Engagement Charts and P-Values

6.7.1 Average Number of Shares

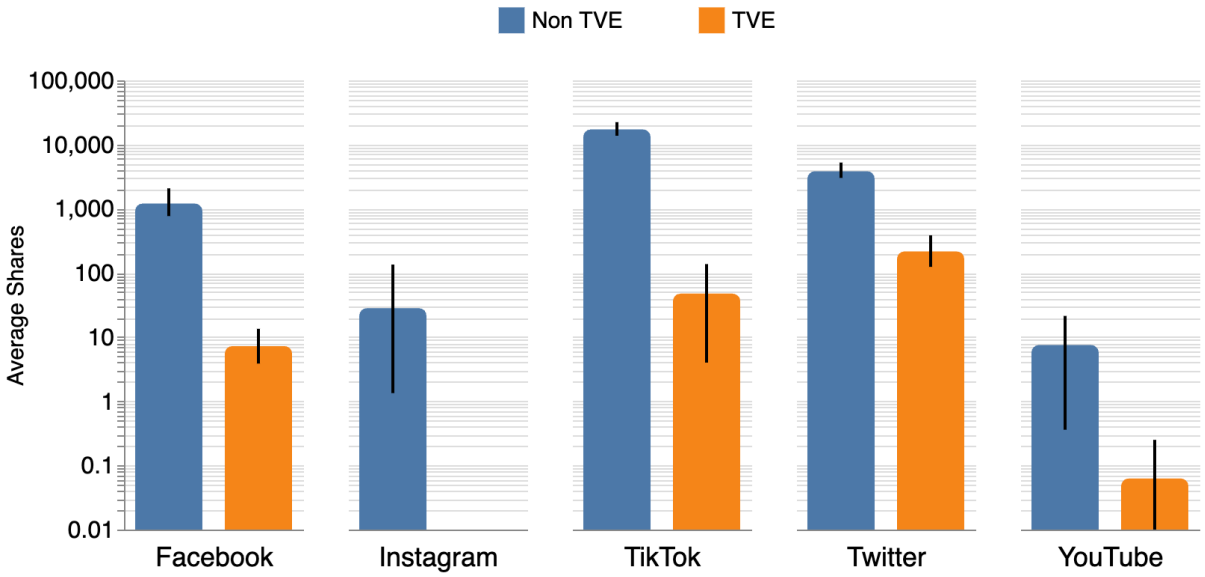

Figure A.6.7.1

This chart shows the average number of shares for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the Interaction charts all use a log scale instead of a linear scale.[29]

[29] Since the range of values in these charts is very high, it would be difficult to visualise the bars that have lower values. Taking the logarithm makes the values closer in range on the log scale and hence easier to visualise the chart as a whole. Therefore, for the interaction charts, we are using log scales.

6.7.2 Average Number of Likes

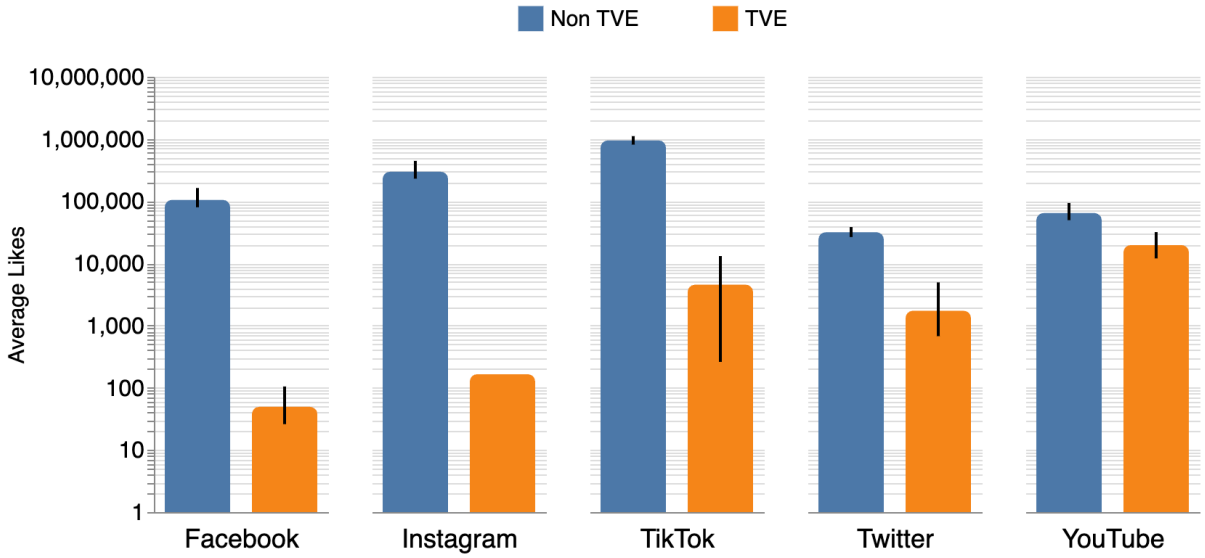

Figure A.6.7.2

This chart shows the average number of likes for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the Interaction charts all use a log scale instead of a linear scale.

6.7.3 Average Number of Comments

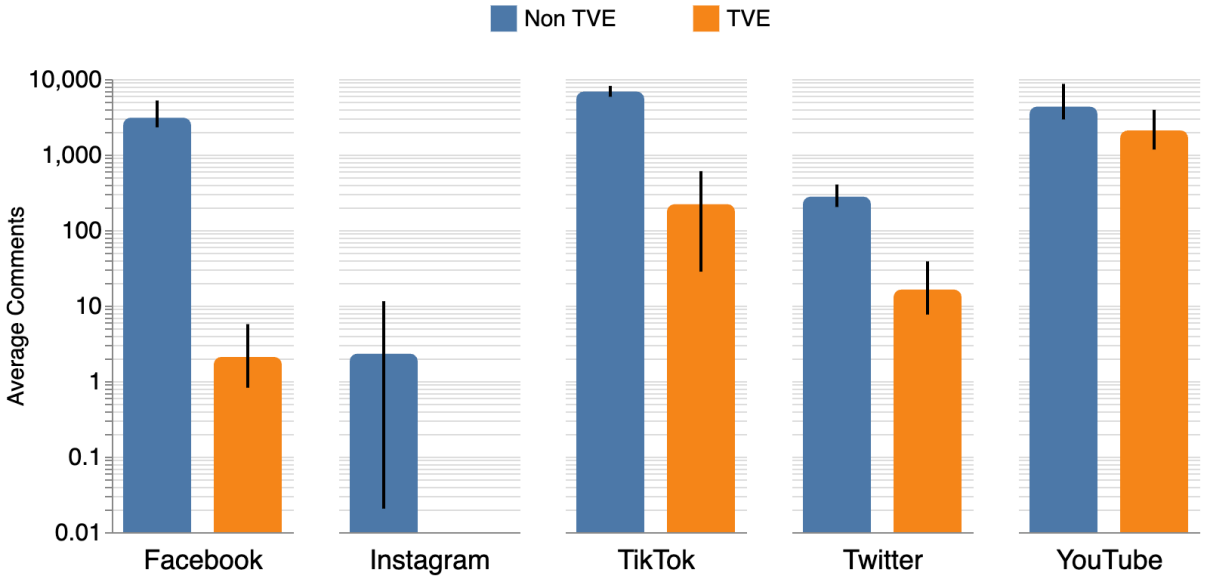

Figure A.6.7.3

This chart shows the average number of comments for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the Interaction charts all use a log scale instead of a linear scale.

6.7.4 Average Number of Followers

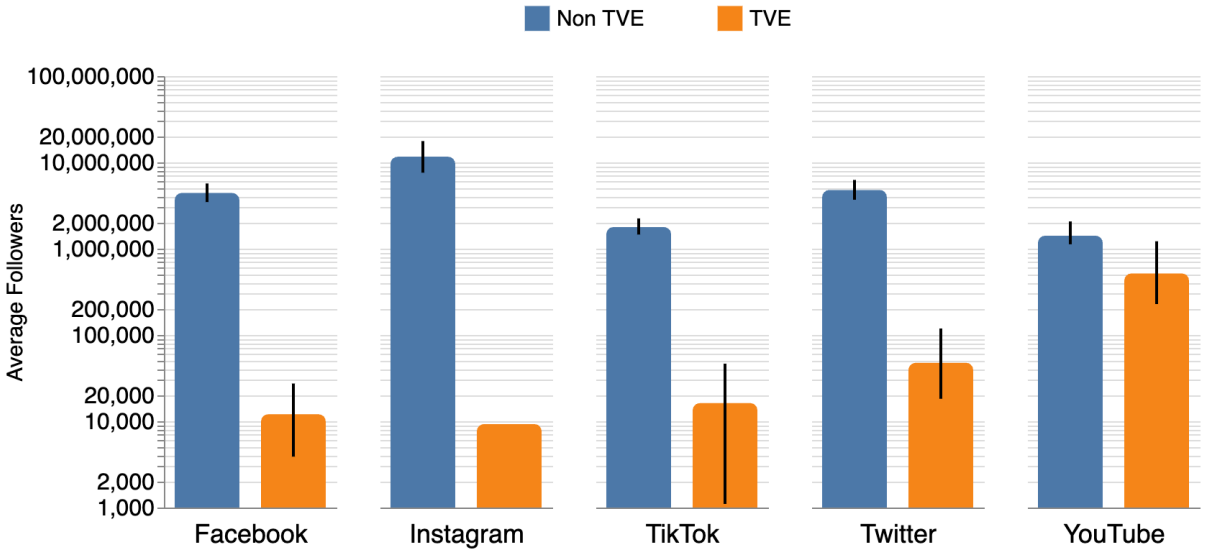

Figure A.6.7.4

This chart shows the average number of followers for TVE and non-TVE content on different platforms. The black lines in the bars represent 90% confidence intervals. Note that the Interaction charts all use a log scale instead of a linear scale.

6.8 Task 1: Removal Rates for TVE Content

6.8.1 Removal Rates per Platform

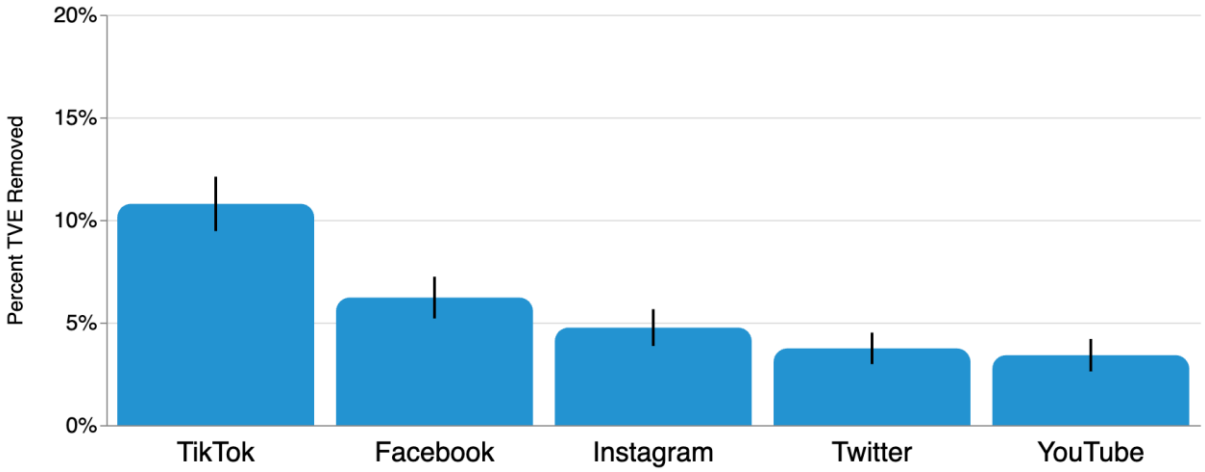

Figure A.6.8.1

This graph shows the percentage of TVE content that was removed on a given platform. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0311	0.2628	0.0000	0.7646
Twitter	0.0497	0.3870	0.0000	
TikTok	0.0062	0.0002		
Instagram	0.2802			

Table A.6.8.1

P-values on Removal Rates per Platform

alpha = 0.01 (Bonferroni correction from 0.05)

6.8.2 Removal Rates per Language

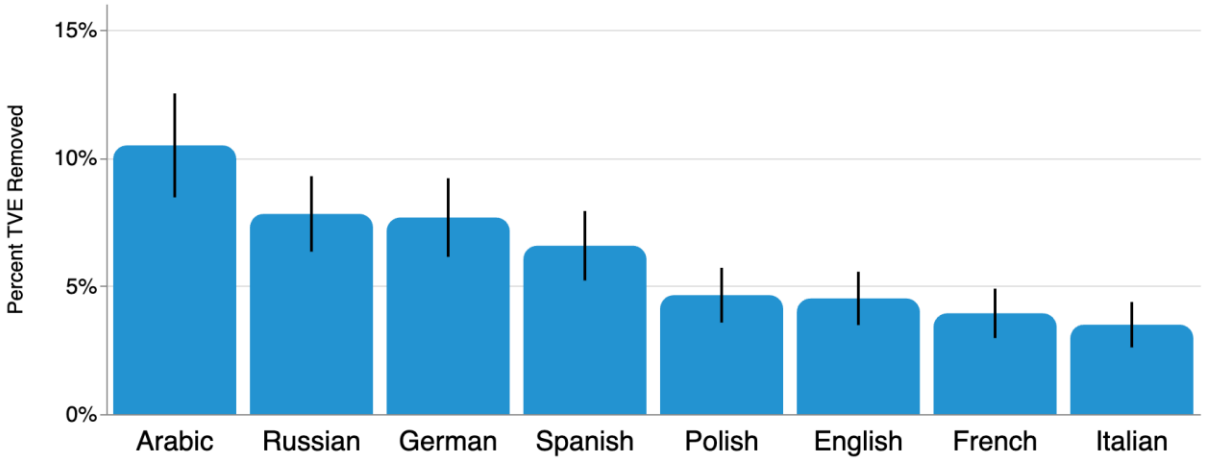

Figure A.6.8.2

This graph shows the percentage of TVE content that was removed in a given language. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0957	0.2222	0.1053	0.5905	0.0490	0.2588	0.5354
Russian	0.2750	0.0618	0.0234	0.9472	0.0086	0.0766	
Polish	0.0053	0.9317	0.6224	0.0965	0.4042		
Italian	0.0003	0.4522	0.7349	0.0124			
German	0.2606	0.0791	0.0316				
French	0.0011	0.6825					
English	0.0039						

Table A.6.8.2

P-values on Removal Rates per Language

alpha = 0.00625 (Bonferroni correction from 0.05)

6.8.3 Removal Rates per TVE Type

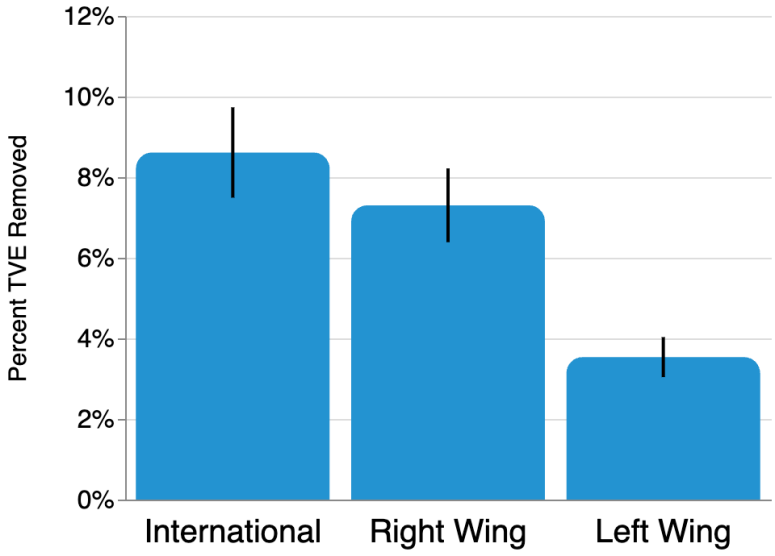

Figure A.6.8.3

This graph shows the percentage of TVE content that was removed per TVE type. The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.3610	0.0001
Left Wing	0.0000	

Table A.6.8.3

P-values on Removal Rates per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.8.4 Average number of Shares associated with Removed TVE

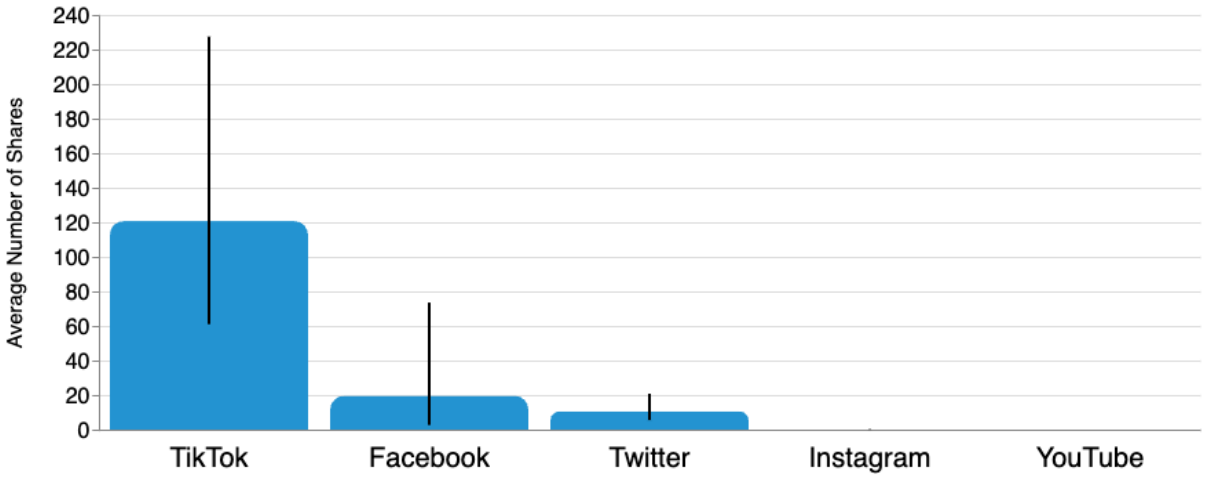

Figure A.6.8.4

This is the average number of shares associated with TVE content that was eventually removed within the period of the study. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.3076	0.2936	0.1426	0.0386
Twitter	0.5656	0.0160	0.1352	
TikTok	0.1001	0.0781		
Instagram	0.2784			

Table A.6.8.4

P-values on Average number of Shares associated with Removed TVE

alpha = 0.01 (Bonferroni correction from 0.05)

6.8.5 Average number of Shares associated with TVE that wasn't Removed

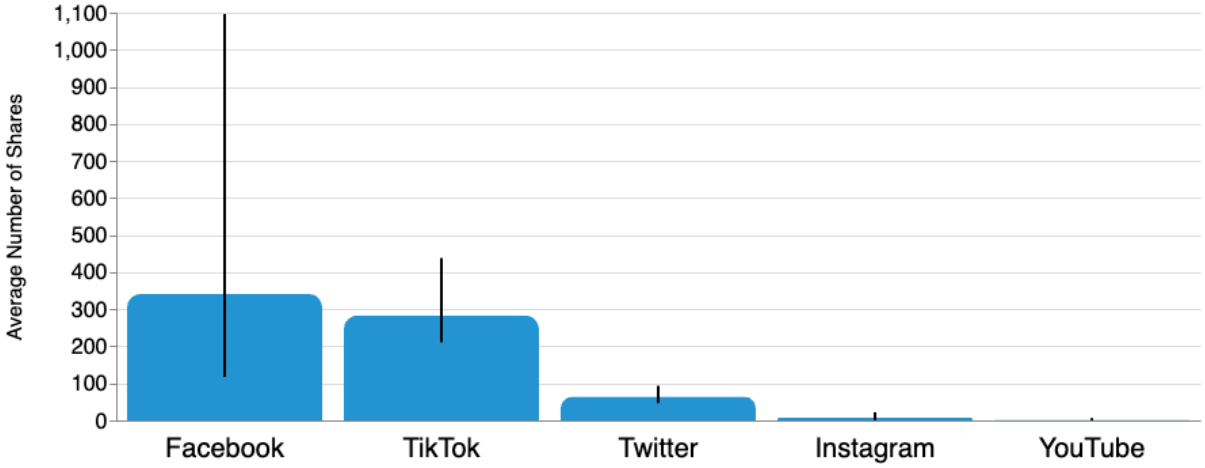

Figure A.6.8.5

This is the average number of shares associated with TVE content that was not removed during the monitoring period of the study. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.1245	0.3010	0.0000	0.0000
Twitter	0.1781	0.0001	0.0001	
TikTok	0.8029	0.0000		
Instagram	0.1198			

Table A.6.8.5

P-values on Average number of Shares associated with non-Removed TVE

alpha = 0.01 (Bonferroni correction from 0.05)

6.8.6 Average number of Shares associated with TVE that wasn't Removed broken down by language

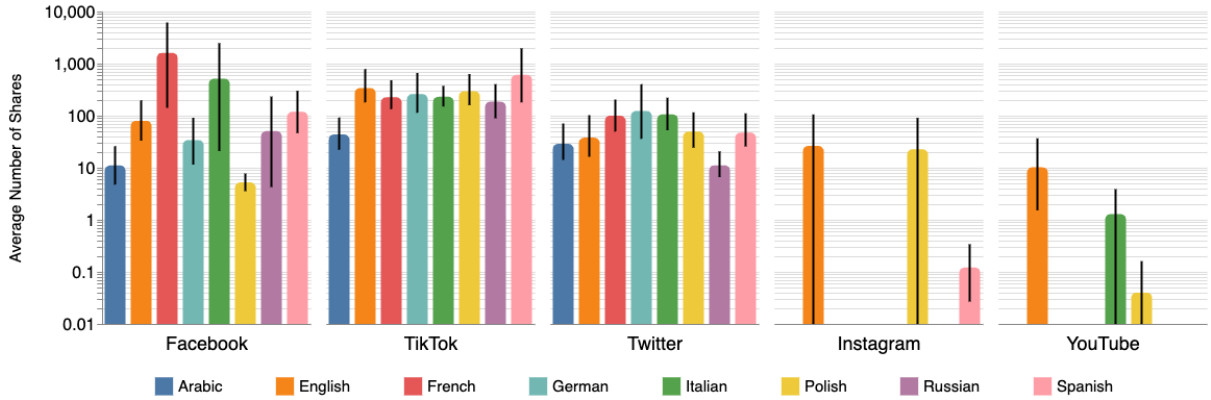

Figure A.6.8.6

This is the average number of shares associated with TVE content that was not removed during the monitoring period of the study broken down by platform and language. The black lines in the bars represent 90% confidence intervals.

6.8.7 Average number of Shares associated with Removed Borderline content

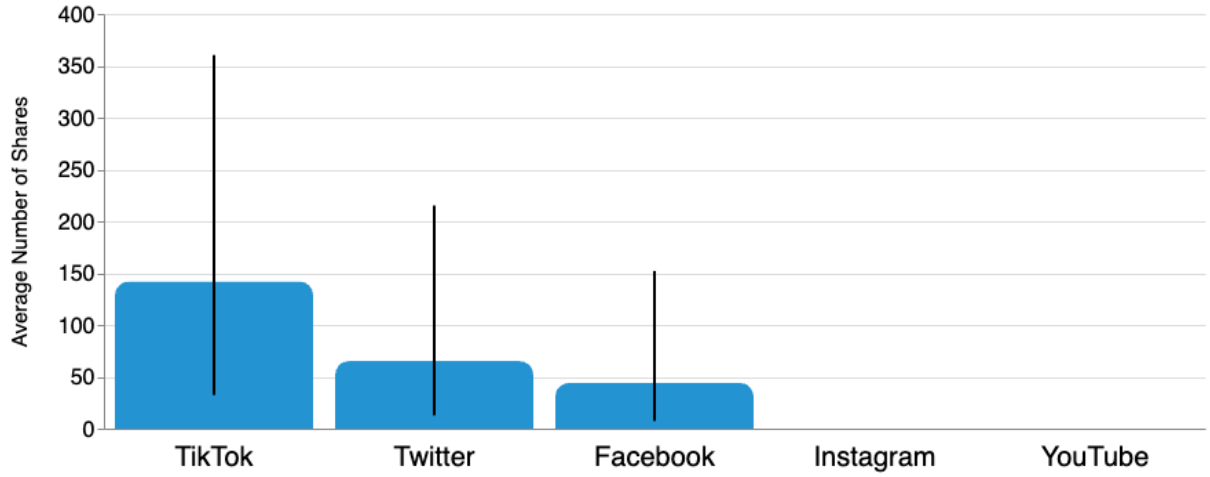

Figure A.6.8.7

This is the average number of shares associated with Borderline content that was eventually removed within the period of the study. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.1305	0.0000	0.1387	0.1215
Twitter	0.6911	0.1194	0.4868	
TikTok	0.3310	0.1567		
Instagram	0.1596			

Table A.6.8.7

P-values on Average number of Shares associated with Removed Borderline content
alpha = 0.01 (Bonferroni correction from 0.05)

6.8.8 Average number of Shares associated with Borderline content that wasn't Removed

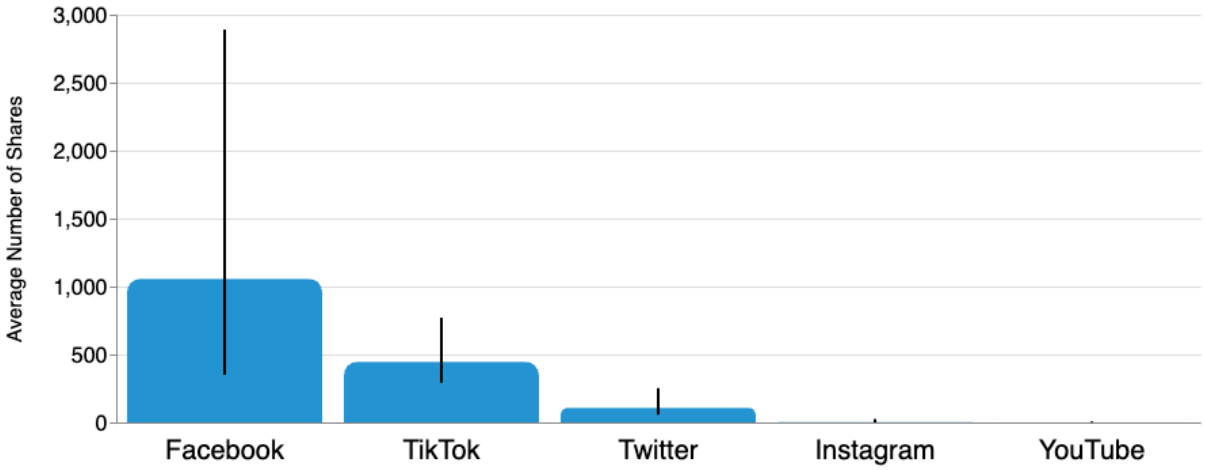

Figure A.6.8.8

This is the average number of shares associated with Borderline content that was not removed during the monitoring period of the study. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0617	0.5061	0.0001	0.0149
Twitter	0.1141	0.0302	0.0083	
TikTok	0.3682	0.0002		
Instagram	0.0861			

Table A.6.8.8

P-values on Average number of Shares associated with non-Removed Borderline content

alpha = 0.01 (Bonferroni correction from 0.05)

6.8.9 Average number of Shares associated with Borderline content that wasn't Removed broken down by language

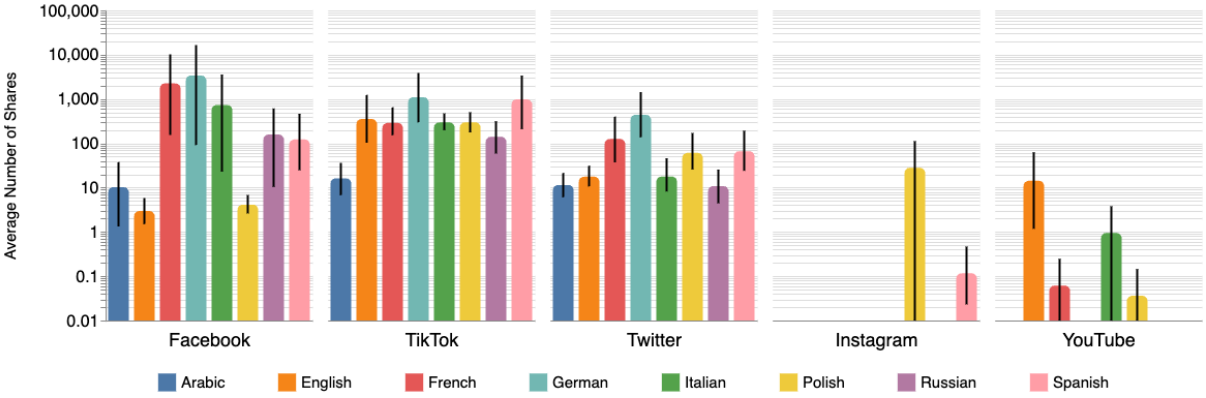

Figure A.6.8.9

This is the average number of shares associated with Borderline content that was not removed during the monitoring period of the study broken down by platform and language. The black lines in the bars represent 90% confidence intervals.

6.9 Task 1: Removal Time

6.9.1 Removal Time per Platform

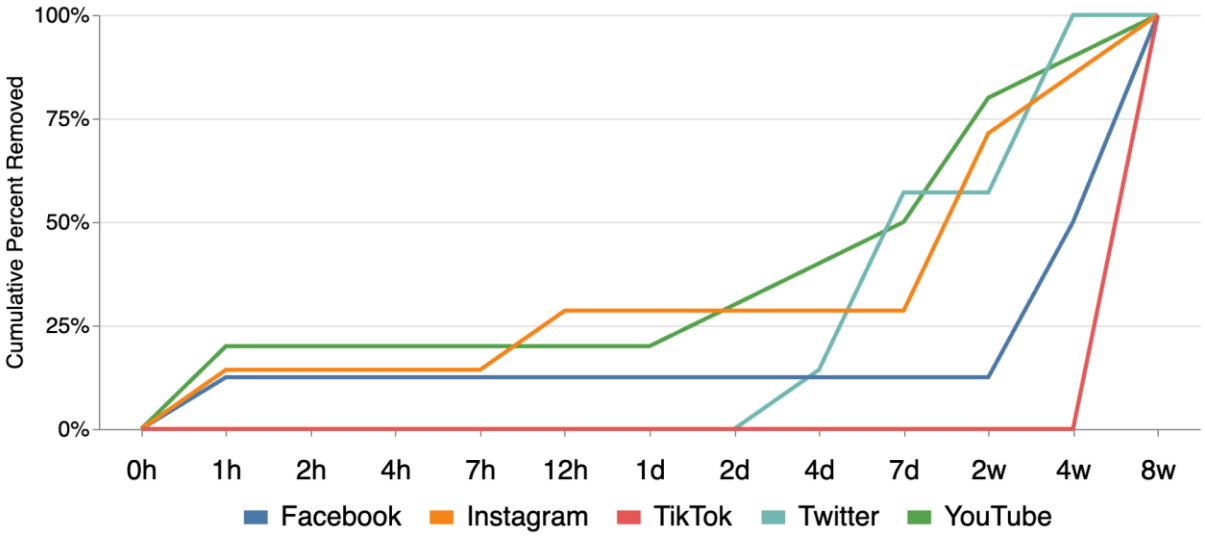

Figure A.6.9.1

This chart shows how long it took for each platform to remove the total amount of TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any TVE content that wasn't removed.

6.9.2 Removal Time per Language

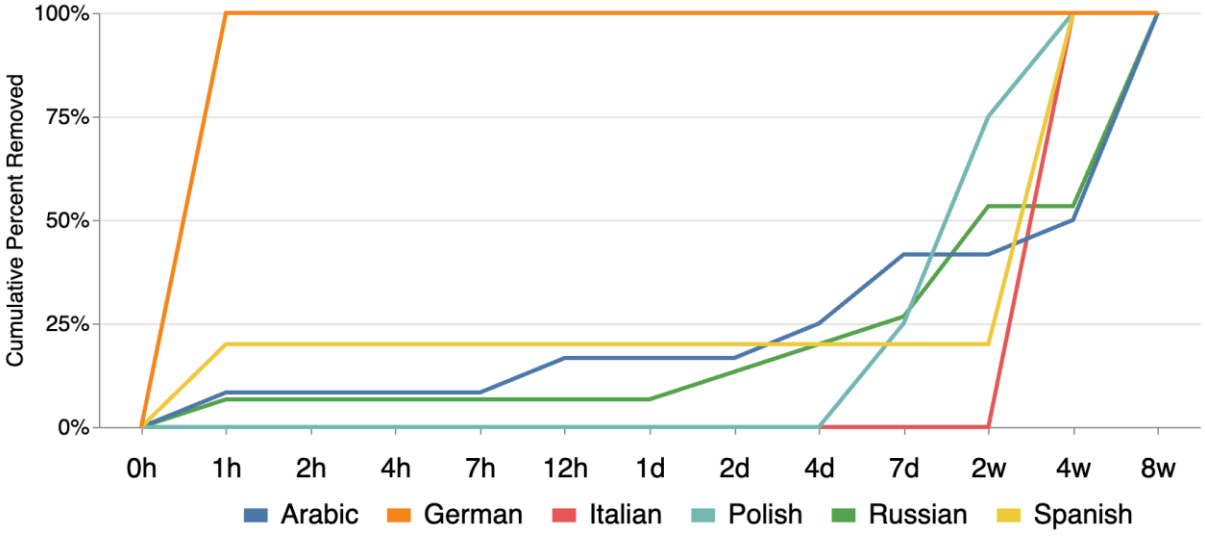

Figure A.6.9.2

This chart shows how long it took for each language to remove the total amount of TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any TVE content that wasn't removed.

6.9.3 Removal Time per TVE Type

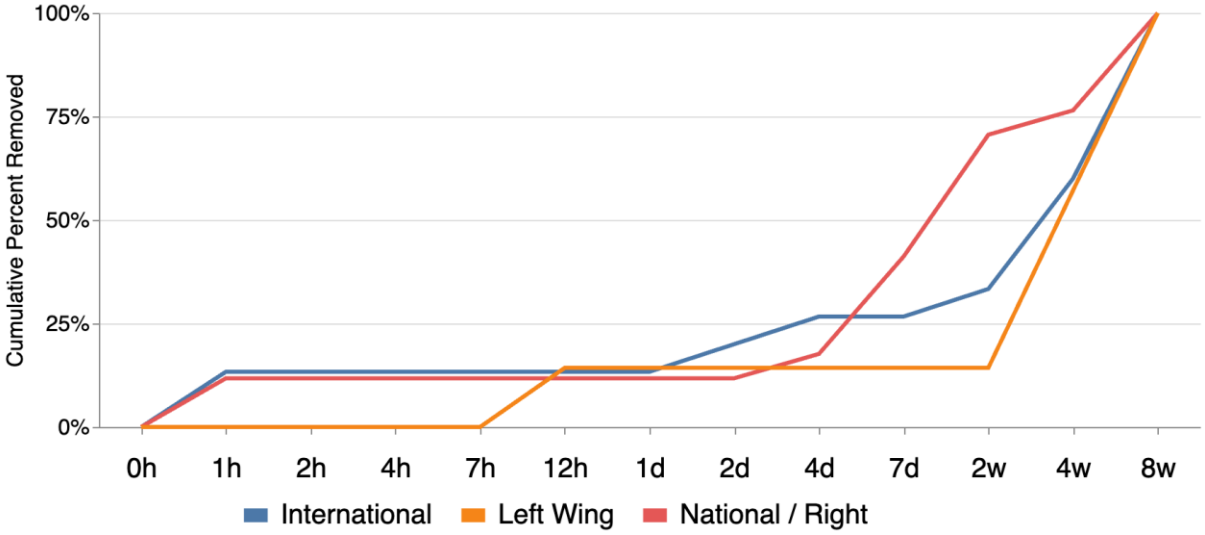

Figure A.6.9.3

This chart shows how long it took for each TVE Type to remove the total amount of TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any TVE content that wasn't removed.

6.10 Task 1: User Sentiment Metrics

6.10.1 Severity Ratings by Platform

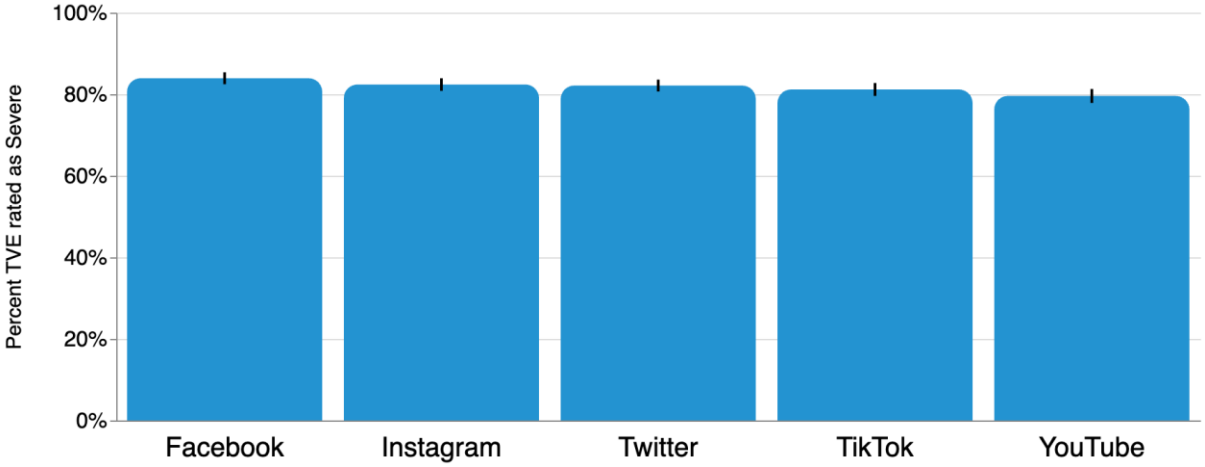

Figure A.6.10.1

This chart shows the percentage of TVE content present on different platforms that was rated by users as "severe". The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0522	0.2201	0.4963	0.2536
Twitter	0.3872	0.9027	0.6548	
TikTok	0.2019	0.5803		
Instagram	0.4703			

Table A.6.10.1

P-values on Severity Ratings per Platform

alpha = 0.01 (Bonferroni correction from 0.05)

6.10.2 Severity Ratings by Language

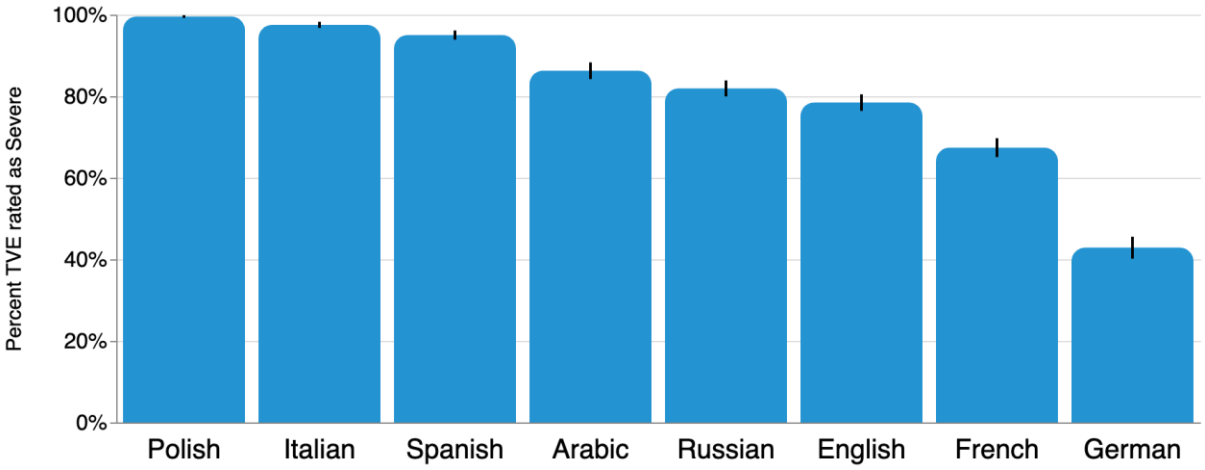

Figure A.6.10.2

This chart shows the percentage of TVE content present in different languages that was rated by users as severe. The black lines in the bars represent 90% confidence intervals.

Polish values are biased here because of the presence of bad actors in the dataset, which caused users to rate everything in the dataset as "severe". This problem was investigated and deemed not present for the other languages.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0001	0.0000	0.0000	0.0000	0.0573	0.0000	0.0000
Russian	0.1311	0.2206	0.0000	0.0000	0.0000	0.0000	
Polish	0.0000	0.0000	0.0000	0.0000	0.0124		
Italian	0.0000	0.0000	0.0000	0.0000			
German	0.0000	0.0000	0.0000				
French	0.0000	0.0003					
English	0.0088						

Table A.6.10.2

P-values on Severity Ratings per Language

alpha = 0.00625 (Bonferroni correction from 0.05)

6.10.3 Severity Ratings by TVE Type

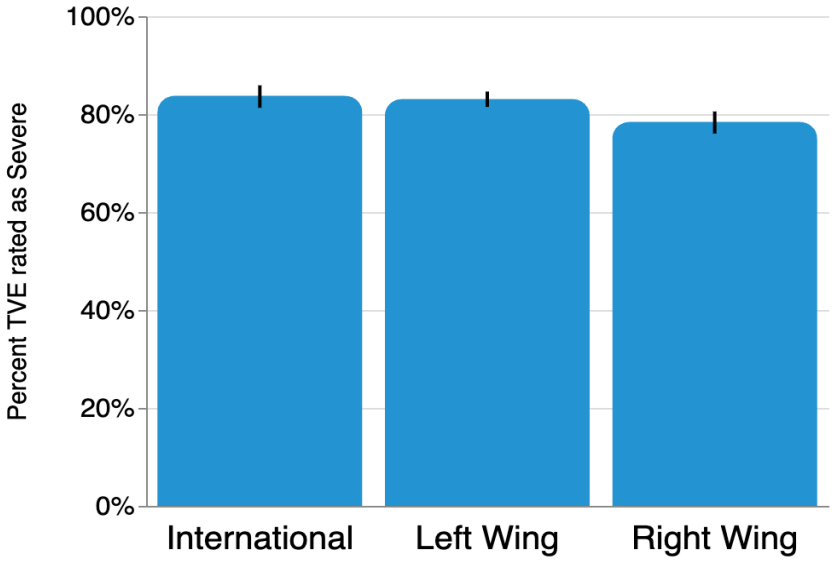

Figure A.6.10.3

This chart shows the percentage of TVE content present for different TVE Types that was rated by users as "severe". The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.0074	0.0045
Left Wing	0.6996	

Table A.6.10.3

P-values on Severity Ratings per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.10.4 Severity Ratings over Time per Platform

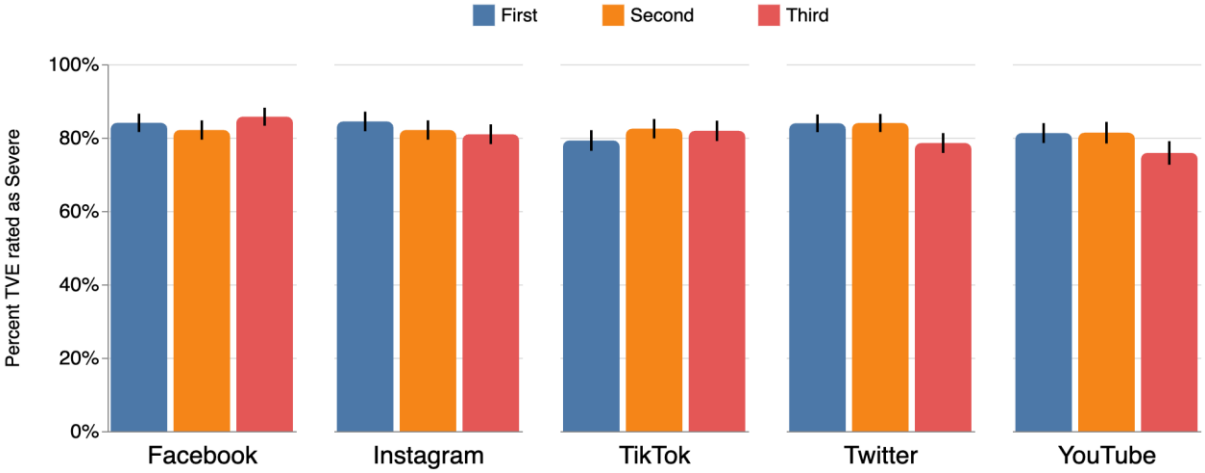

Figure A.6.10.4

This chart shows the percentage of TVE content rated as "Severe". First, second and third stages are shown for each platform. The black lines in the bars represent 90% confidence intervals.

P-values are not included for this chart due to the number of combinations and complexity of the comparison.

6.11 Task 2: Amplification Charts

6.11.1 Amplification Across all Platforms

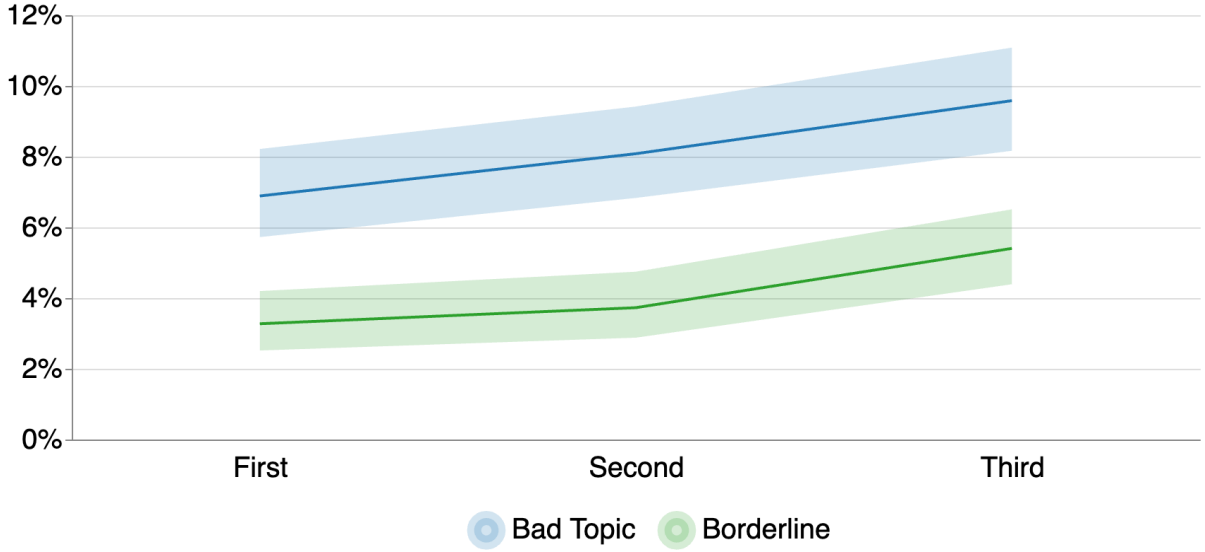

Figure A.6.11.1a

This graph illustrates the mean percentage of "Bad Topic" and "Borderline" content recommended in user feeds over time. The shaded area represents the 90% confidence interval (alpha = 0.05)

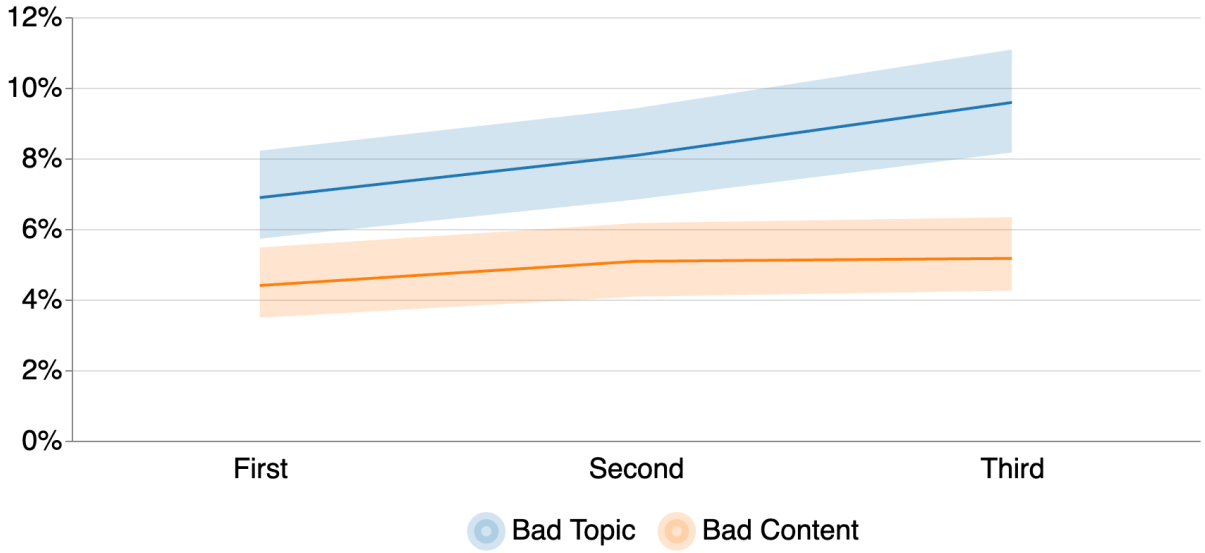

Figure A.6.11.1b

This graph illustrates the mean percentage of "Bad Topic" and "Bad Content" content recommended in user feeds over time. The shaded area represents the 90% confidence interval (alpha = 0.05)

6.11.2 Amplification Average Percent of Bad Content in Feed per Platform

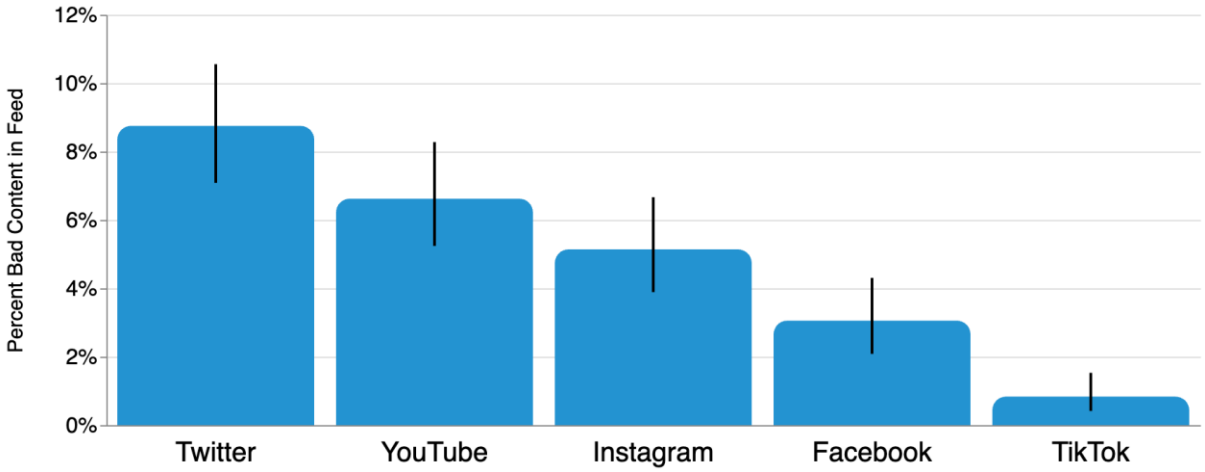

Figure A.6.11.2

This chart shows the average percentage of Bad (TVE) Content in the feeds of respective platforms. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.0522	0.2201	0.4963	0.2536
Twitter	0.3872	0.9027	0.6548	
TikTok	0.2019	0.5803		
Instagram	0.4703			

Table A.6.11.2

P-values on Amplification Average Percent of Bad Content in Feed per Platform

alpha = 0.01 (Bonferroni correction from 0.05)

6.11.3 Amplification Average Percent of Bad Content in Feed per Language

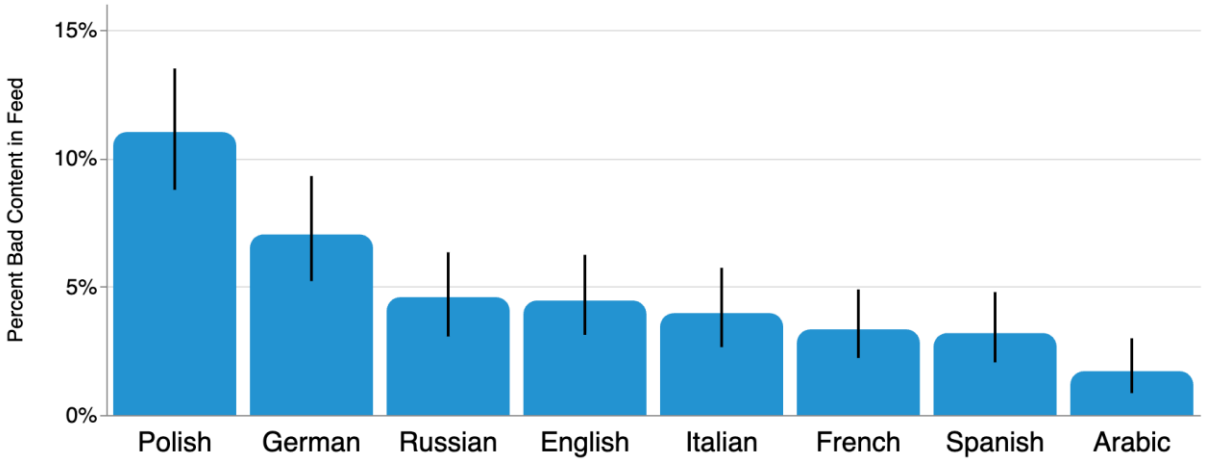

Figure A.6.11.3

This chart shows the average percentage of Bad (TVE) Content in the feeds of respective languages. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.1466	0.3265	0.9038	0.0097	0.5335	0.0000	0.2809
Russian	0.0118	0.9246	0.3343	0.1168	0.6438	0.0003	
Polish	0.0000	0.0002	0.0000	0.0387	0.0001		
Italian	0.0381	0.7145	0.6132	0.0440			
German	0.0001	0.0986	0.0126				
French	0.1146	0.3855					
English	0.0155						

Table A.6.11.3

P-values on Amplification Average Percentage of Bad Content in Feed per Language

alpha = 0.00625 (Bonferroni correction from 0.05)

6.11.4 Amplification Average Percent of Bad Content in Feed per TVE Type

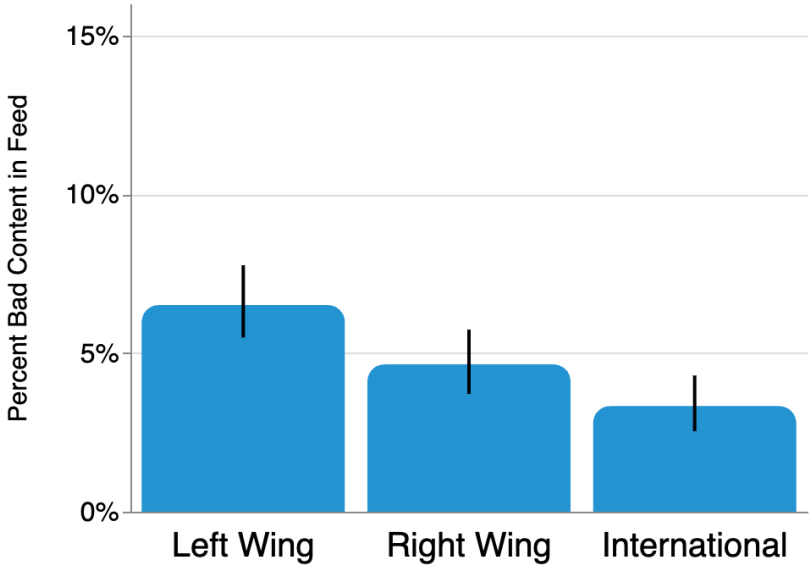

Figure A.6.11.4

This chart shows the average percentage of Bad (TVE) Content in the feeds of respective TVE Types. The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.1058	0.0442
Left Wing	0.0003	

Table A.6.11.4

P-values on Amplification Average Percent of Bad Content in Feed per TVE Type

alpha = 0.017 (Bonferroni correction from 0.05)

6.11.5 Amplification Percentage Change for Bad Content per Content Type

To better compare data across other dimensions (platform, language, etc), we will use the percent difference between the first and third stage's mean. We don't compare to zero state because the assumption is that a new account's feed will be empty or contain zero TVE related content.

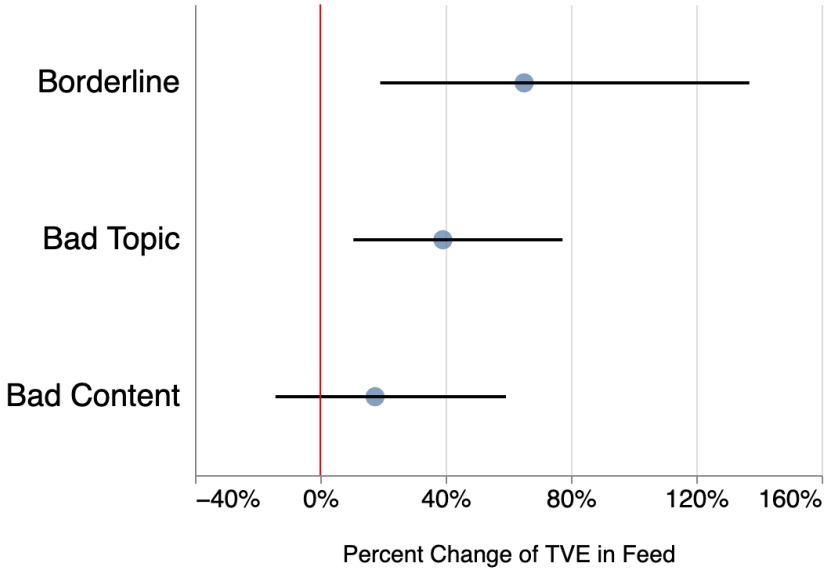

Figure A.6.11.5

This chart shows the percentage change for the amount of Bad (TVE) Content per Content Type in the platform's feeds from the First Stage to the Third Evaluation Phase. The black line represents the 90% confidence intervals. The red line represents the threshold for whether a filter bubble is increasing or decreasing in size.

6.11.6 Amplification Percent Change for Bad Content per Platform

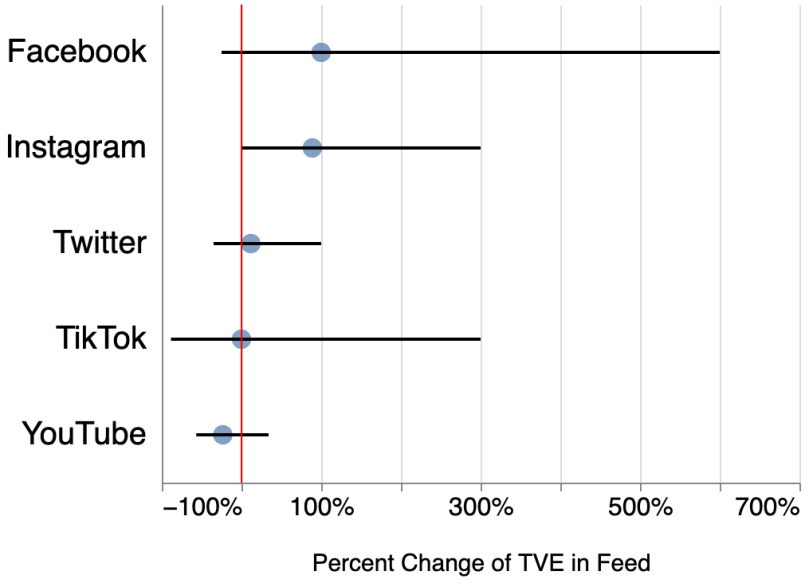

Figure A.6.11.6

This chart shows the percentage change for the amount of Bad (TVE) Content in the platform's feeds from the First Stage to the Third Evaluation Phase. The black line represents the 90% confidence intervals. The red line represents the threshold for whether a filter bubble is increasing or decreasing in size.

6.11.7 Amplification Percent Change for Bad Content per Language

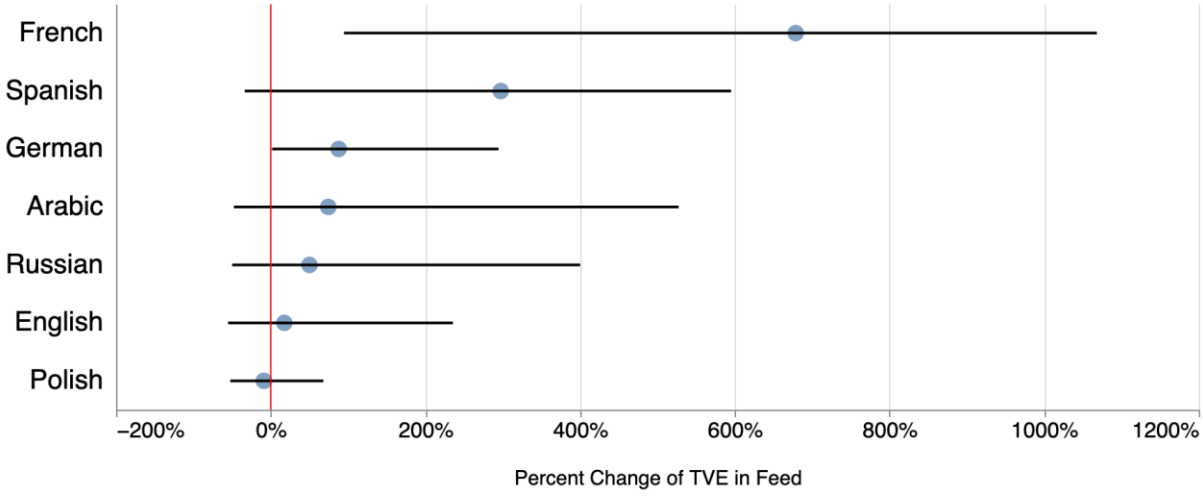

Figure A.6.11.7

This chart shows the percentage change for the amount of Bad (TVE) Content per language from the First Stage to the Third Evaluation Phase. The black line represents the 90% confidence intervals. The red line represents the threshold for whether a filter bubble is increasing or decreasing in size.

Italian is missing from the above chart because there did not exist any Bad Content in feeds for the first search stage within the sample.

6.11.8 Amplification per Platform

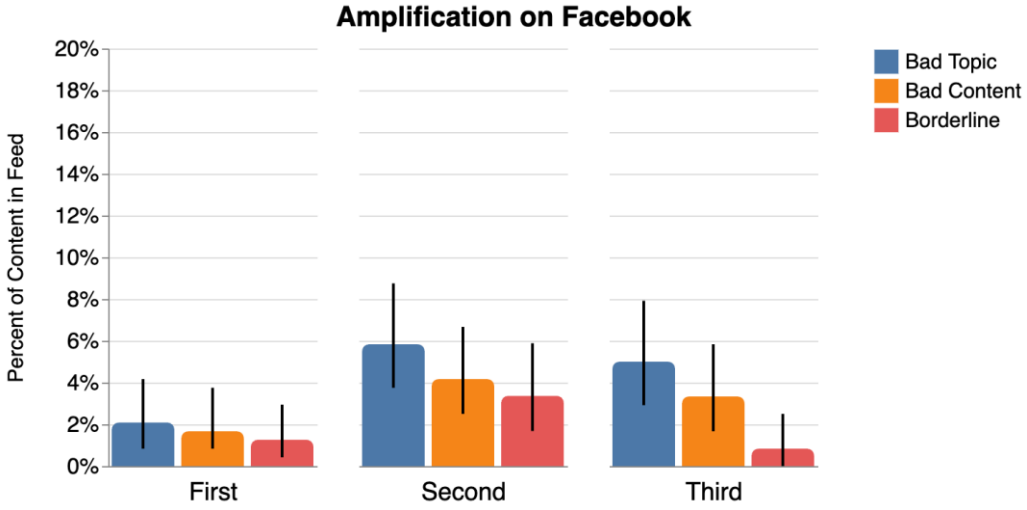

Figure A.6.11.8a

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on Facebook for each stage. The black lines in the bars represent 90% confidence intervals.

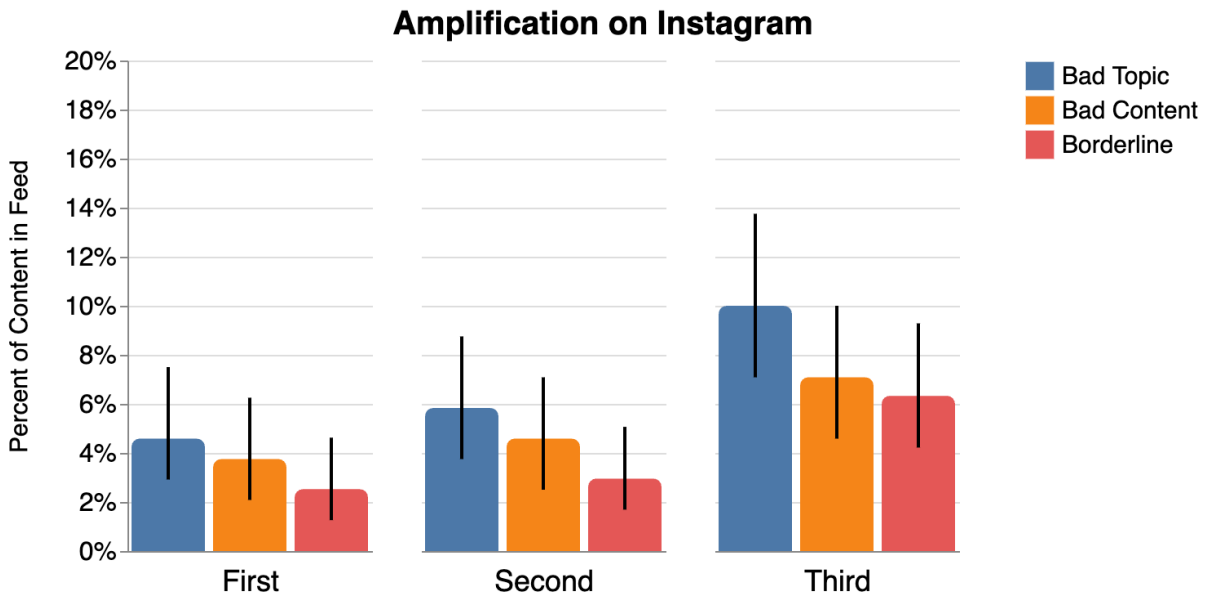

Figure A.6.11.8b

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on Instagram for each stage. The black lines in the bars represent 90% confidence intervals.

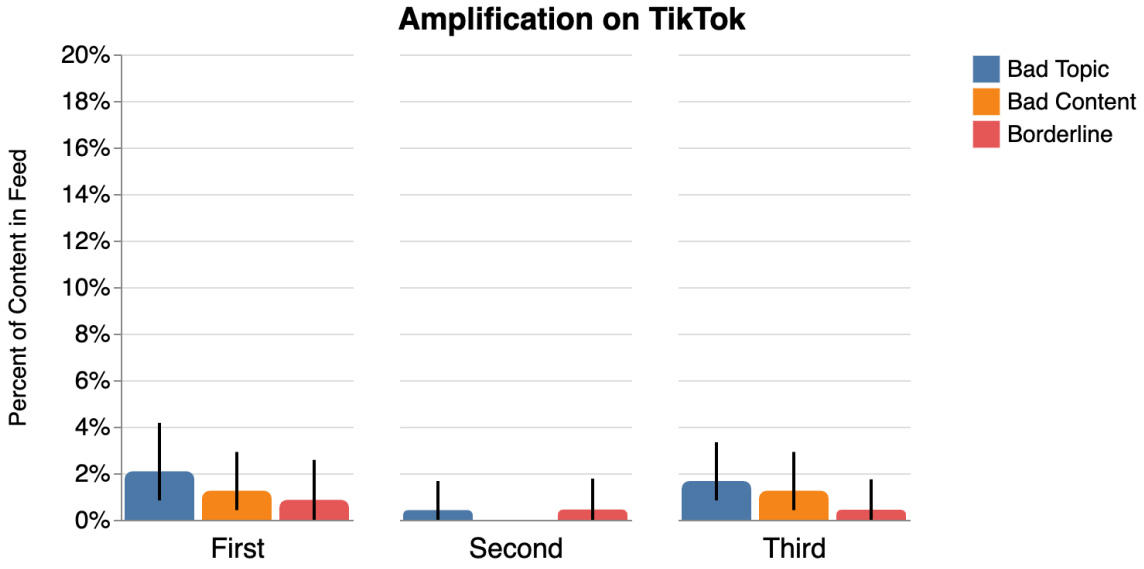

Figure A.6.11.8c

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on TikTok for each stage. The black lines in the bars represent 90% confidence intervals.

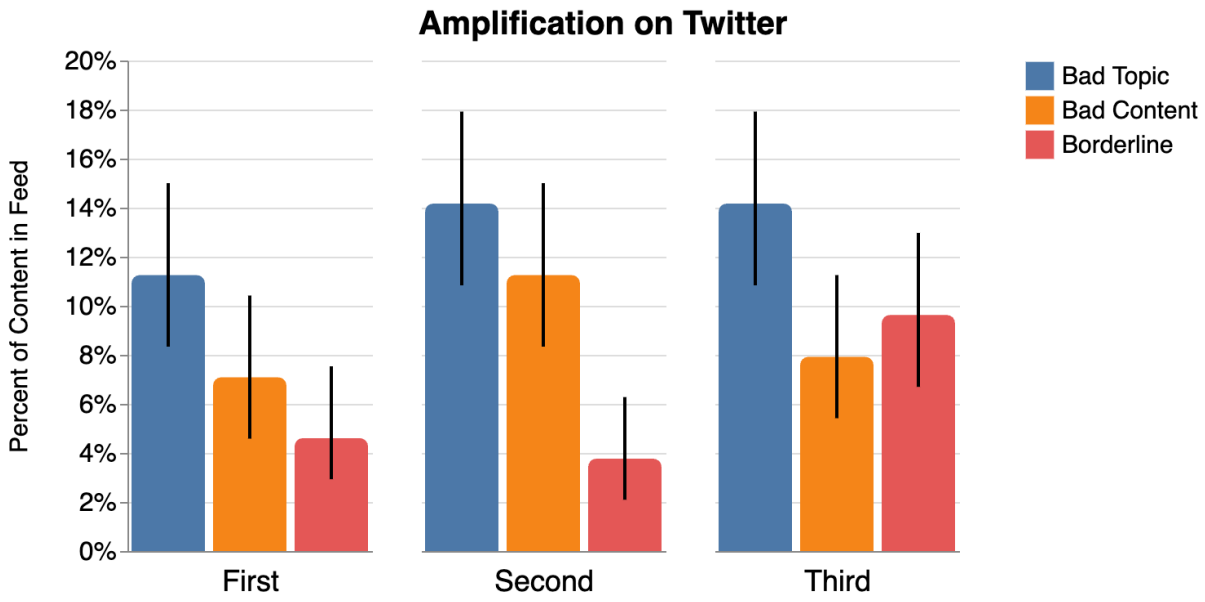

Figure A.6.11.8d

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on Twitter for each stage. The black lines in the bars represent 90% confidence intervals.

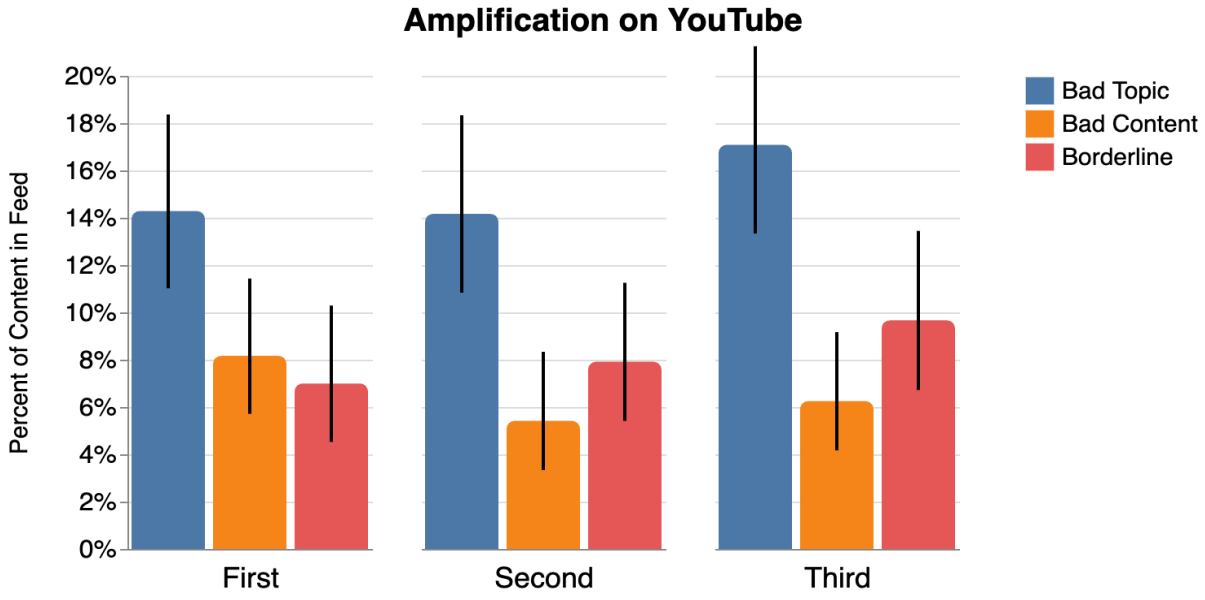

Figure A.6.11.8e

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" on YouTube for each stage. The black lines in the bars represent 90% confidence intervals.

6.11.9 Amplification per Language

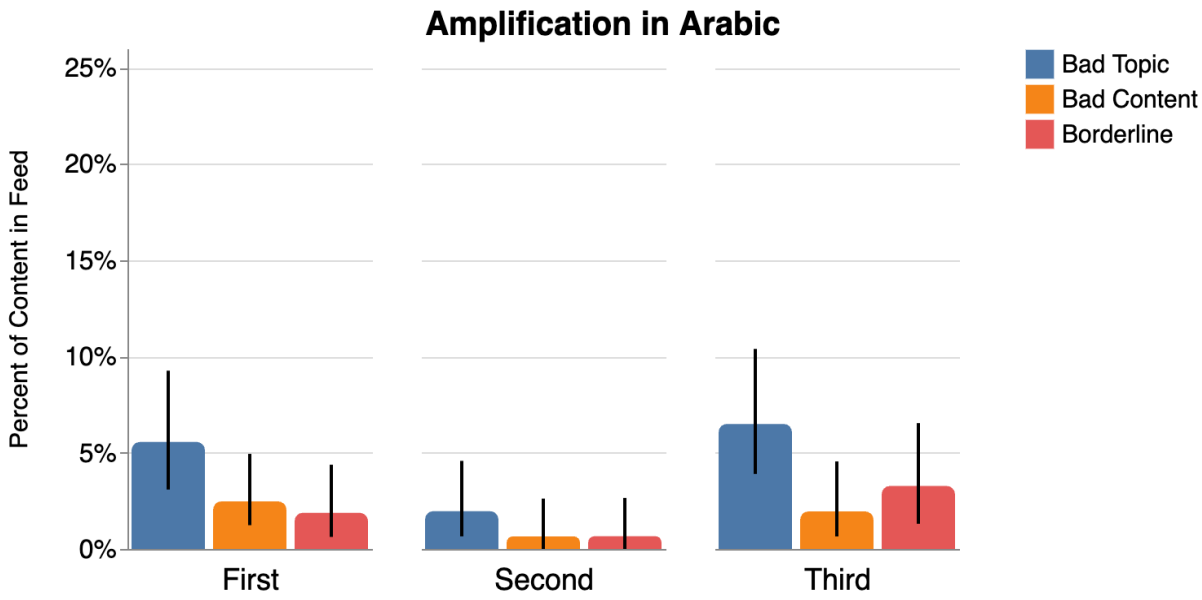

Figure A.6.11.9a

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Arabic for each stage. The black lines in the bars represent 90% confidence intervals.

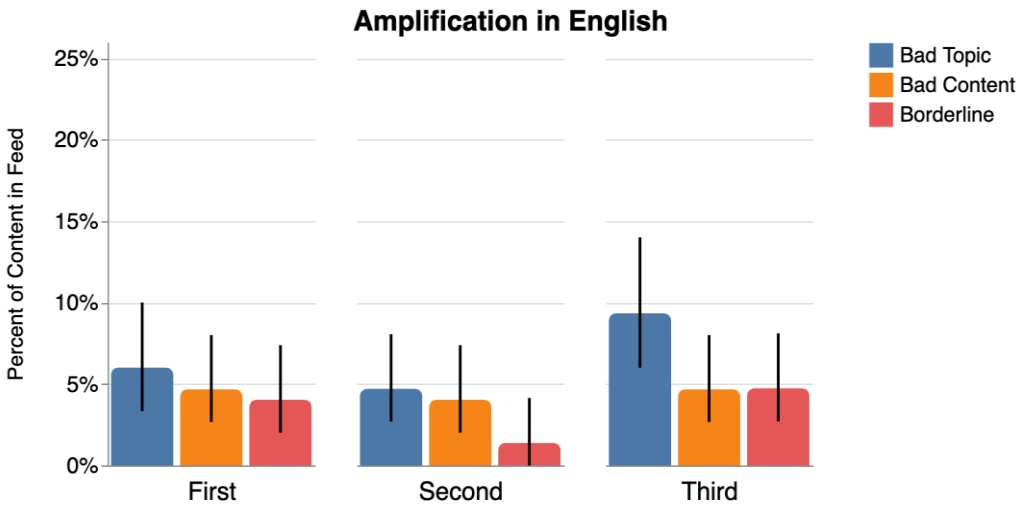

Figure A.6.11.9b

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in English for each stage. The black lines in the bars represent 90% confidence intervals.

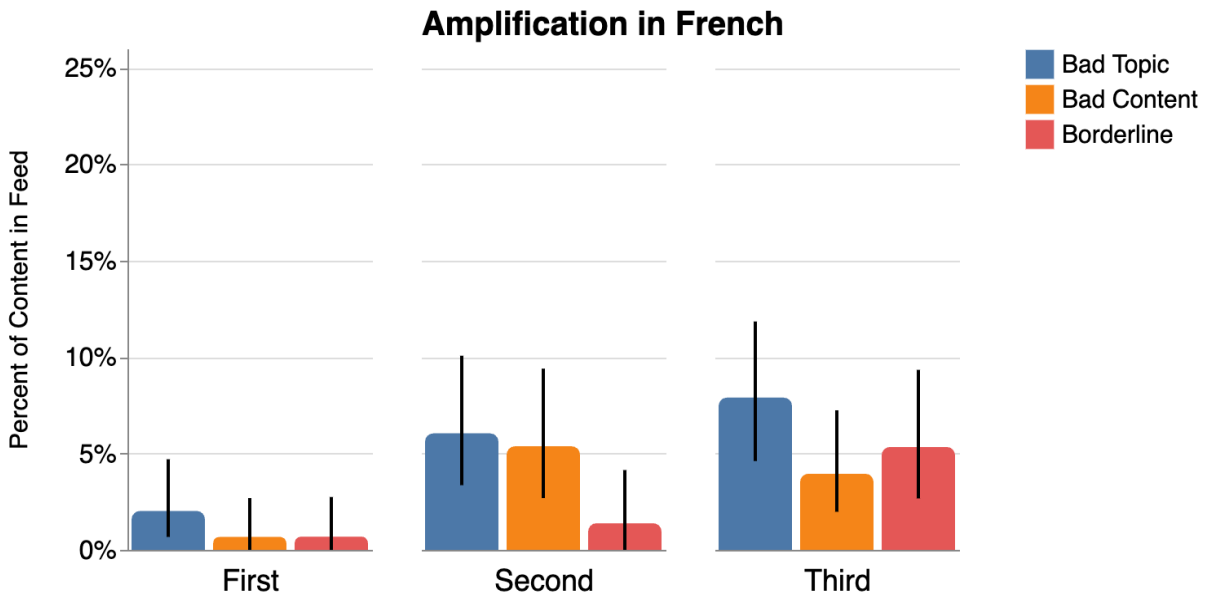

Figure A.6.11.9c

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in French for each stage. The black lines in the bars represent 90% confidence intervals.

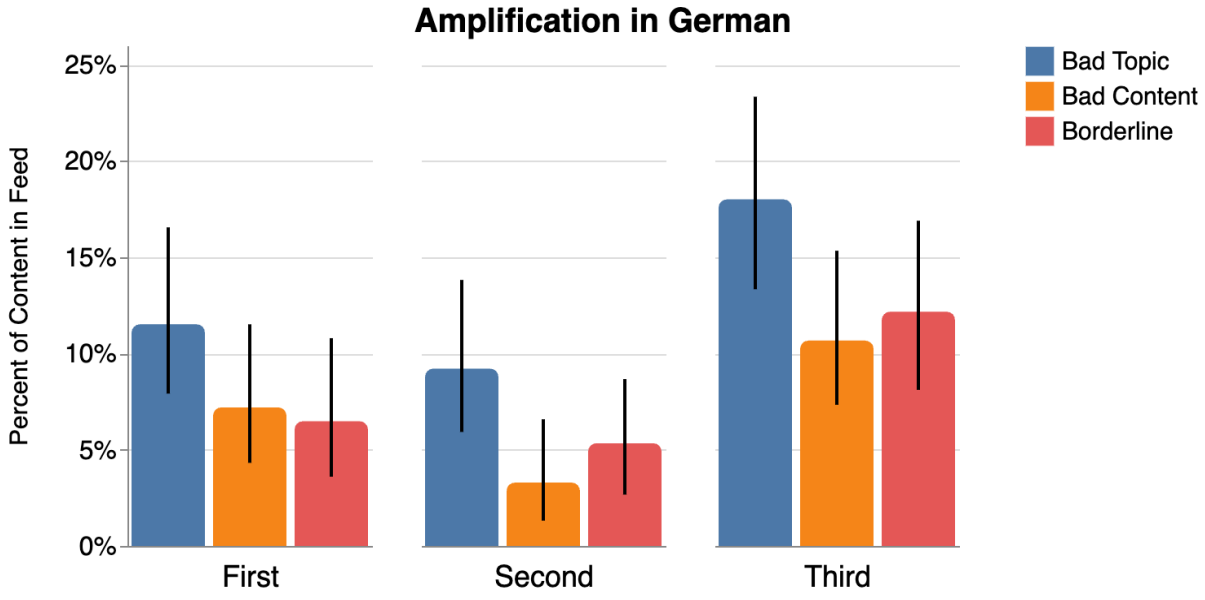

Figure A.6.11.9d

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in German for each stage. The black lines in the bars represent 90% confidence intervals.

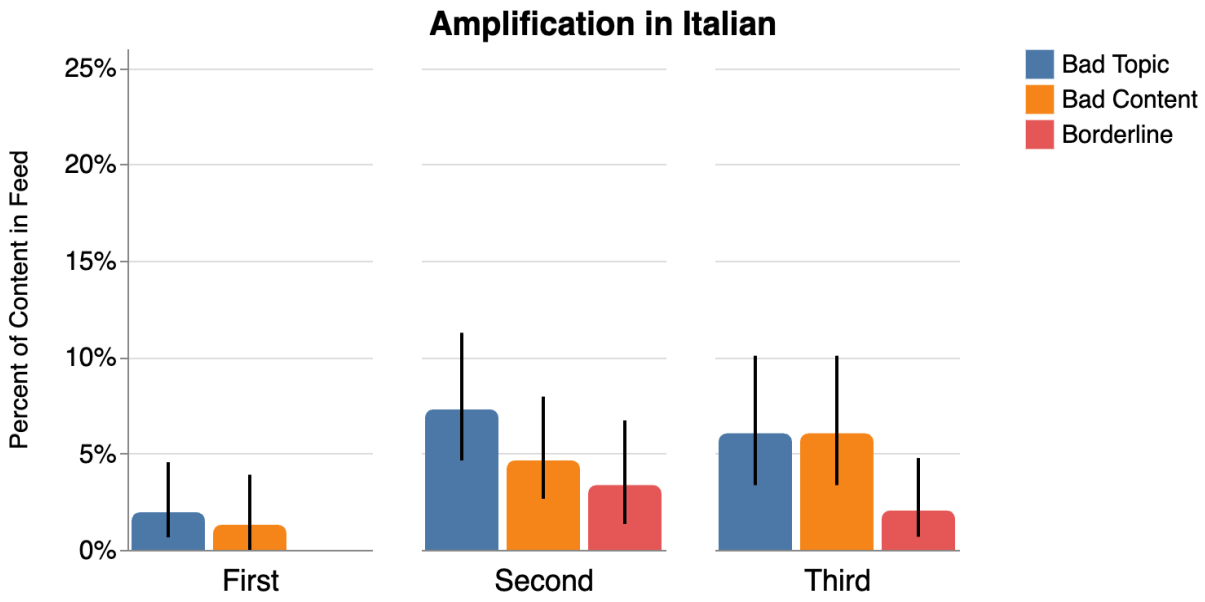

Figure A.6.11.9e

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Italian for each stage. The black lines in the bars represent 90% confidence intervals.

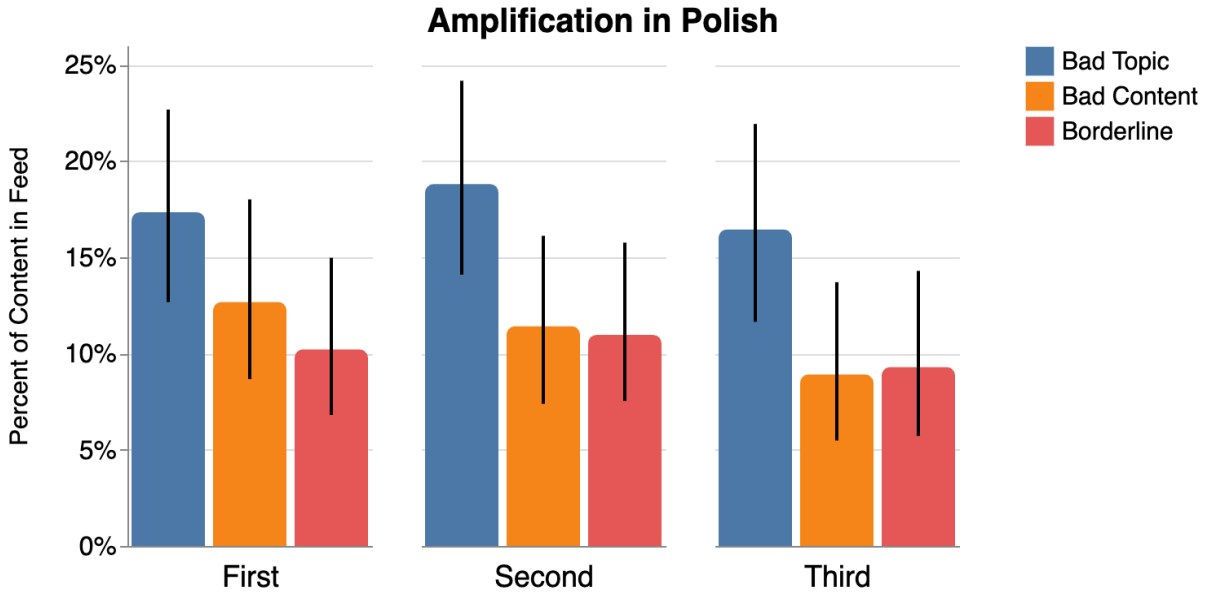

Figure A.6.11.9f

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Polish for each stage. The black lines in the bars represent 90% confidence intervals.

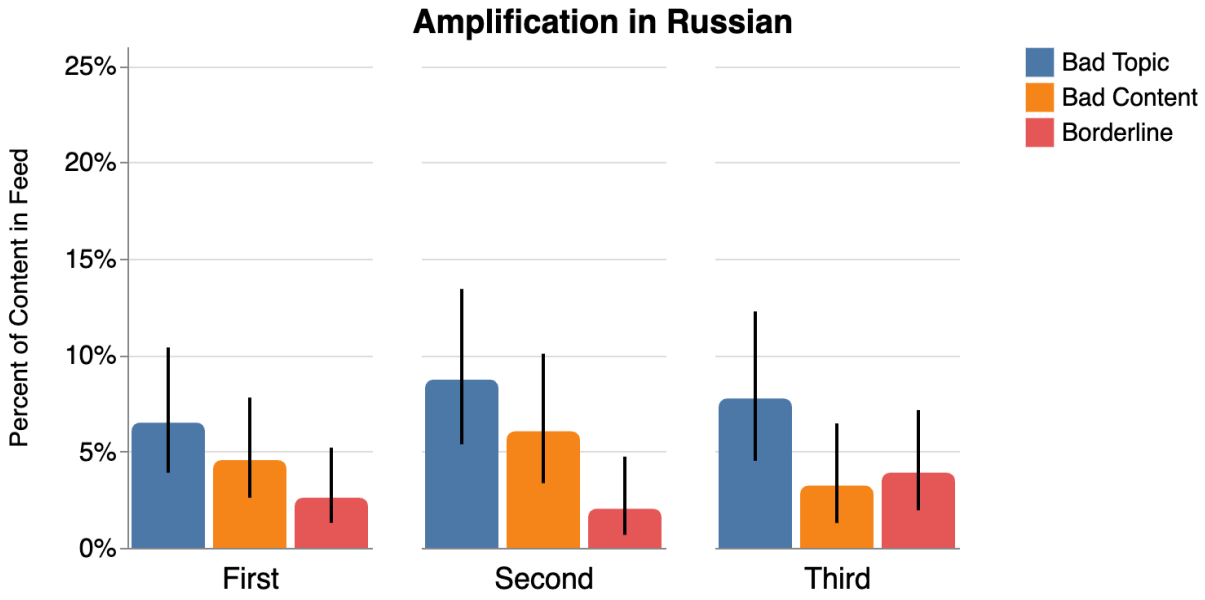

Figure A.6.11.9g

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Russian for each stage. The black lines in the bars represent 90% confidence intervals.

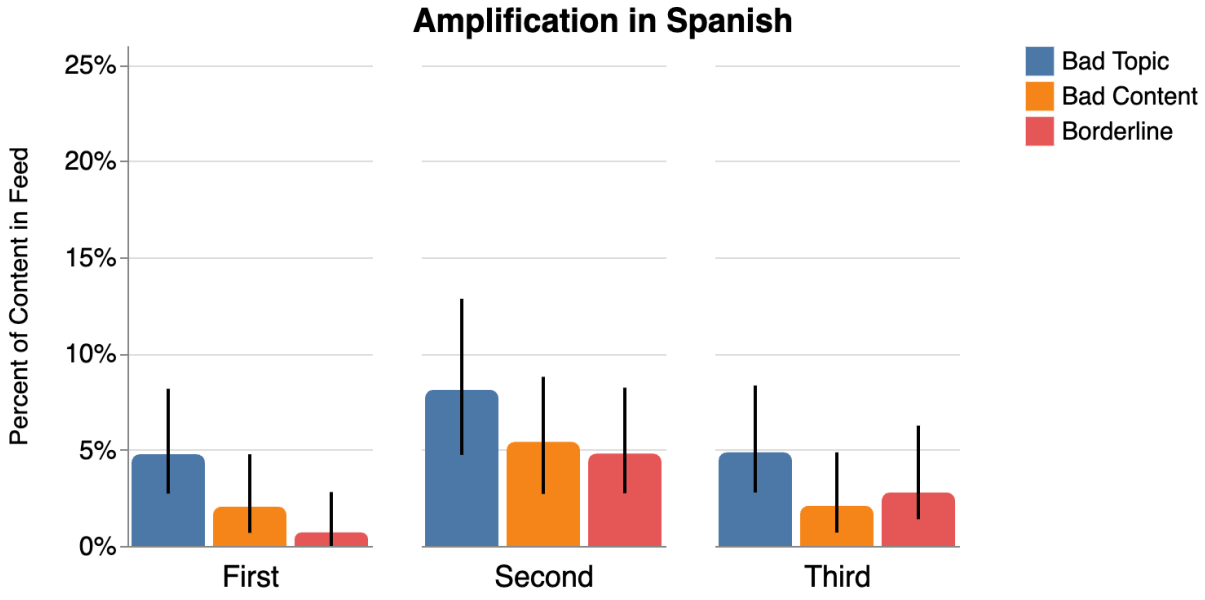

Figure A.6.11.9h

This graph illustrates the percentage of recommended content that is "Bad Content", "Bad Topic", and "Borderline" in Spanish for each stage. The black lines in the bars represent 90% confidence intervals.

6.11.10 Amplification per TVE Type

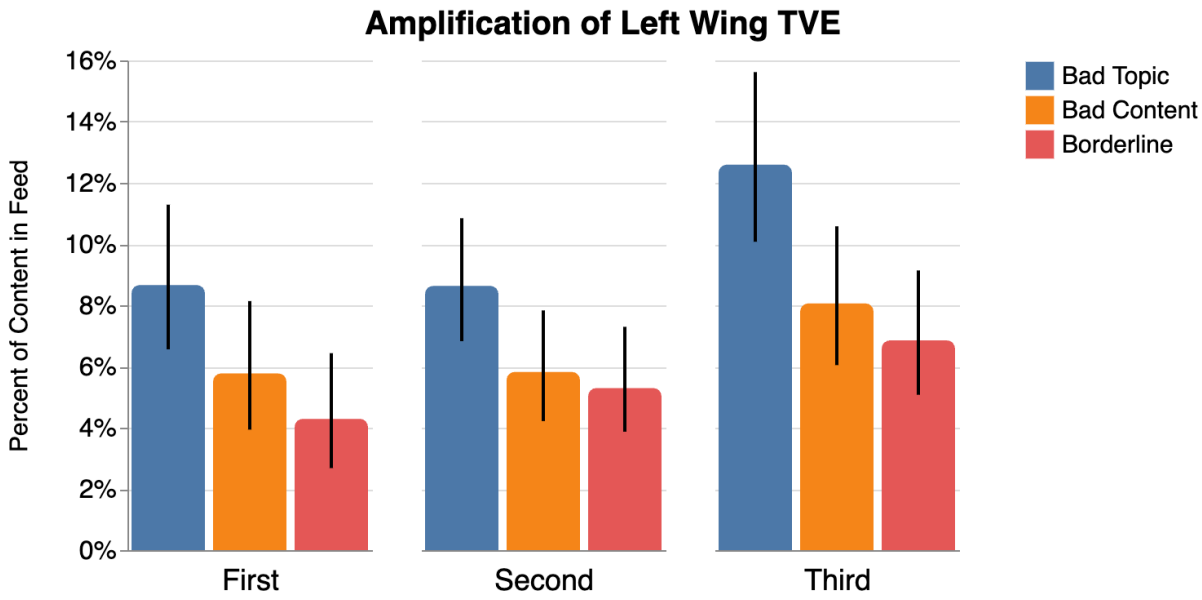

Figure A.6.11.10a

This graph illustrates the percentage of recommended Left-Wing TVE content that is "Bad Content", "Bad Topic", and "Borderline" for each stage. The black lines in the bars represent 90% confidence intervals.

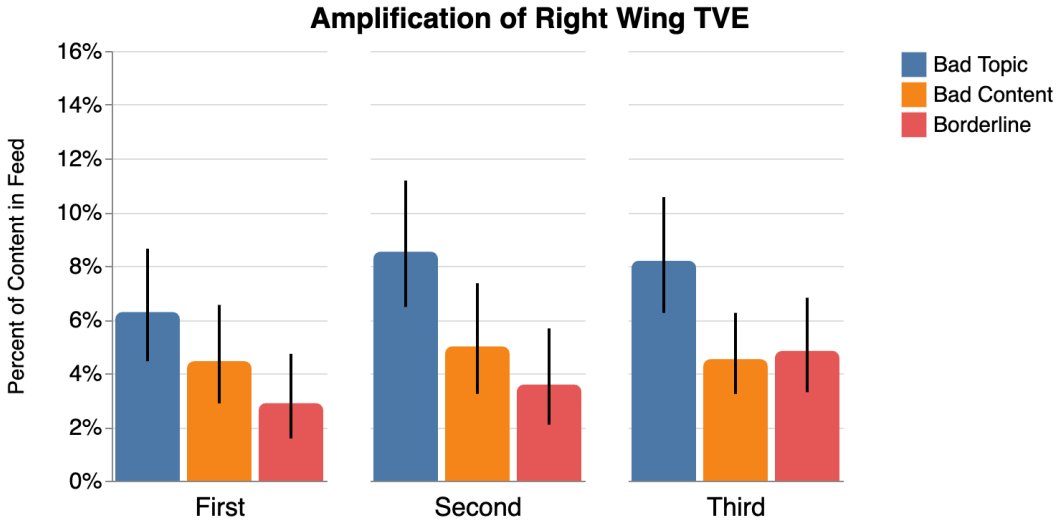

Figure A.6.11.10b

This graph illustrates the percentage of recommended National/Right-Wing TVE content that is "Bad Content", "Bad Topic", and "Borderline" for each stage. The black lines in the bars represent 90% confidence intervals.

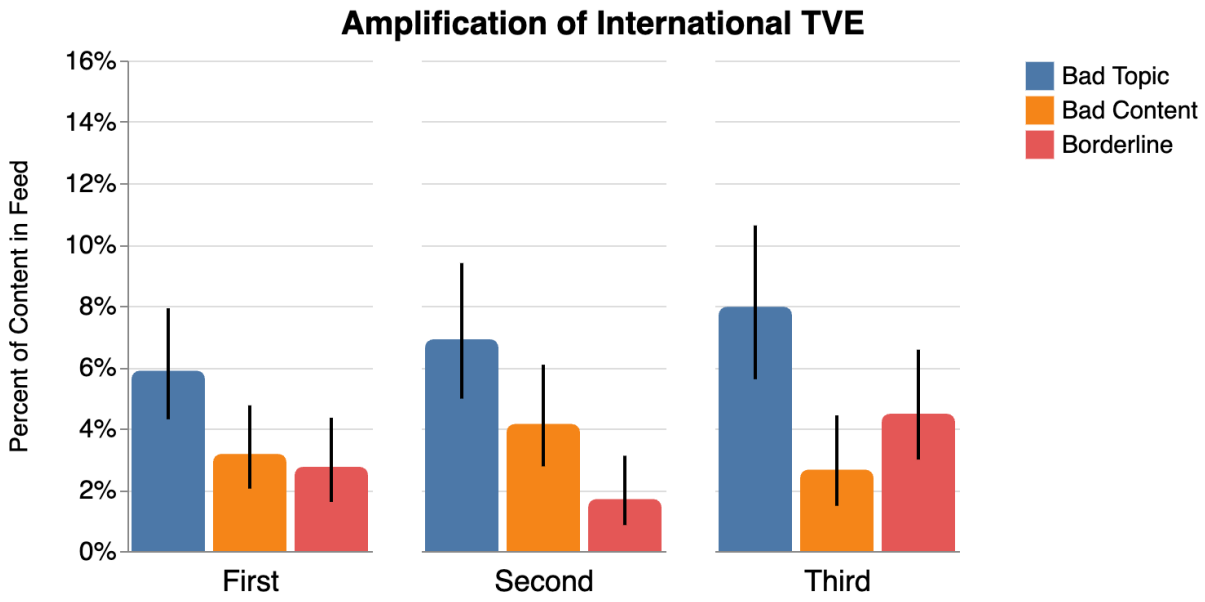

Figure A.6.11.10c

This graph illustrates the percentage of recommended International TVE content that is "Bad Content", "Bad Topic", and "Borderline" for each stage. The black lines in the bars represent 90% confidence intervals.

6.11.11 Amplification P-values

Amplification Per Platform

Platform	Content Type	First/Second	First/Third	Second/Third
['Facebook']	Bad Topic	0.0352	0.0842	0.6875
['Facebook']	Bad Content	0.1041	0.2431	0.6317
['Facebook']	Borderline	0.1278	0.6538	0.0553
['Instagram']	Bad Topic	0.5387	0.0225	0.0913
['Instagram']	Bad Content	0.6486	0.1071	0.2435
['Instagram']	Borderline	0.7791	0.0447	0.0811
['TikTok']	Bad Topic	0.1007	0.7371	0.1782
['TikTok']	Bad Content	0.0826	1.0000	0.0826
['TikTok']	Borderline	0.5634	0.5634	1.0000
['Twitter']	Bad Topic	0.3384	0.3384	1.0000
['Twitter']	Bad Content	0.1142	0.7296	0.2156
['Twitter']	Borderline	0.6486	0.0328	0.0104
['YouTube']	Bad Topic	0.9701	0.3978	0.3799
['YouTube']	Bad Content	0.2306	0.4166	0.6976
['YouTube']	Borderline	0.6820	0.2908	0.5192

P-values on Amplification Average Percent by Content Type in Feed per Platform over time

Amplification per Language

Language	Content Type	First/Second	First/Third	Second/Third
['Arabic']	Bad Topic	0.0963	0.7269	0.0488
['Arabic']	Bad Content	0.1988	0.7540	0.3189
['Arabic']	Borderline	0.3441	0.4317	0.1014

Amplification per TVE Type

tve_type	Content Type	First/Second	First/Third	Second/Third
['International']	Bad Topic	0.5541	0.2520	0.5937
['International']	Bad Content	0.4608	0.6750	0.2793
['International']	Borderline	0.3139	0.1953	0.0317
['Left Wing']	Bad Topic	0.9888	0.0758	0.0538
['Left Wing']	Bad Content	0.9755	0.2102	0.1874
['Left Wing']	Borderline	0.4823	0.1127	0.3203
['National / Right']	Bad Topic	0.2485	0.2901	0.8635
['National / Right']	Bad Content	0.7282	0.9579	0.7547
['National / Right']	Borderline	0.6203	0.1643	0.3995

Table A.6.11.11c

P-values on Amplification Average Percent by Content Type in Feed per TVE Type over time

6.11.12 Interactivity and Amplification

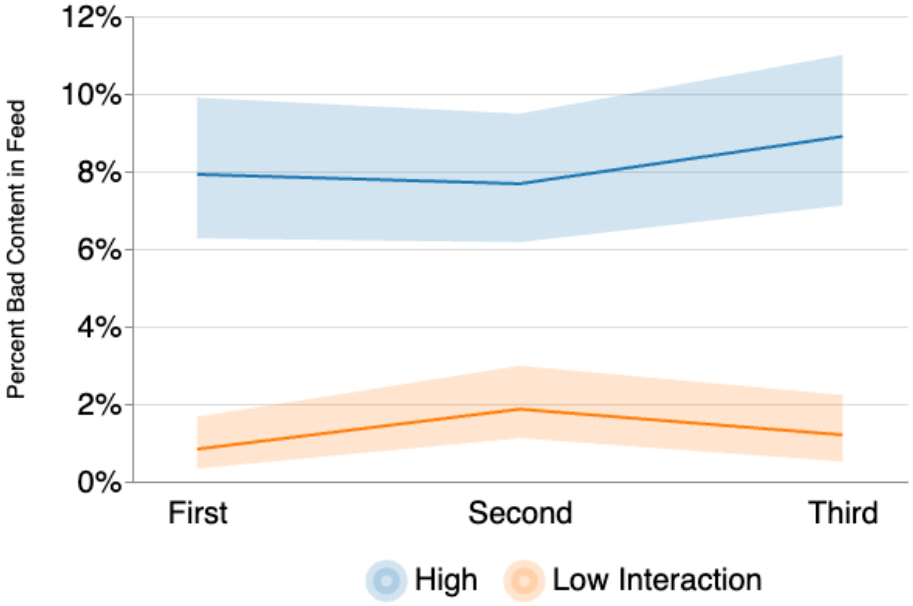

Figure A.6.11.12

This chart shows the average percentage of Bad Content for High and Low interaction over time. The shaded area represents the 90% confidence interval.

6.11.13 Findability and Amplification

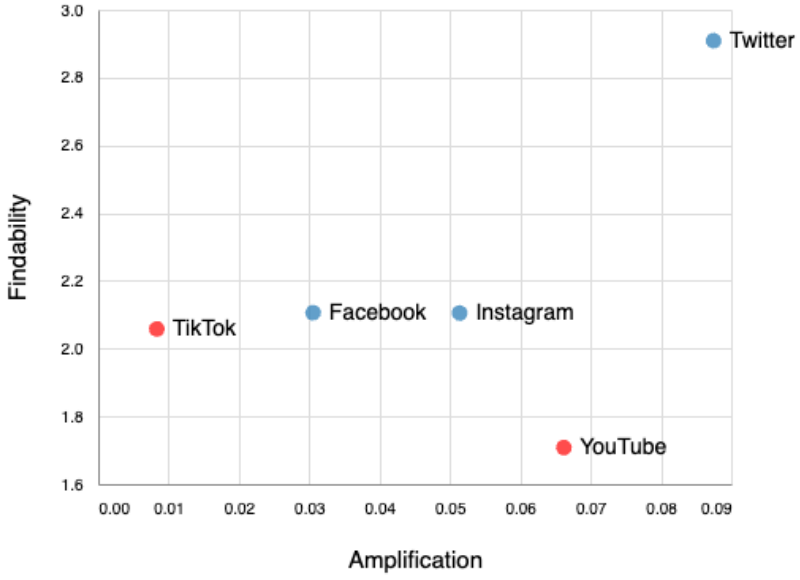

Figure A.6.11.13

This chart shows the comparison of the Findability score and the average percentage of Bad Content in the feed (here called Amplification). TikTok and YouTube are highlighted in red to contrast tech investment in feed vs search.

6.11.14 Engagement Ratio and Amplification

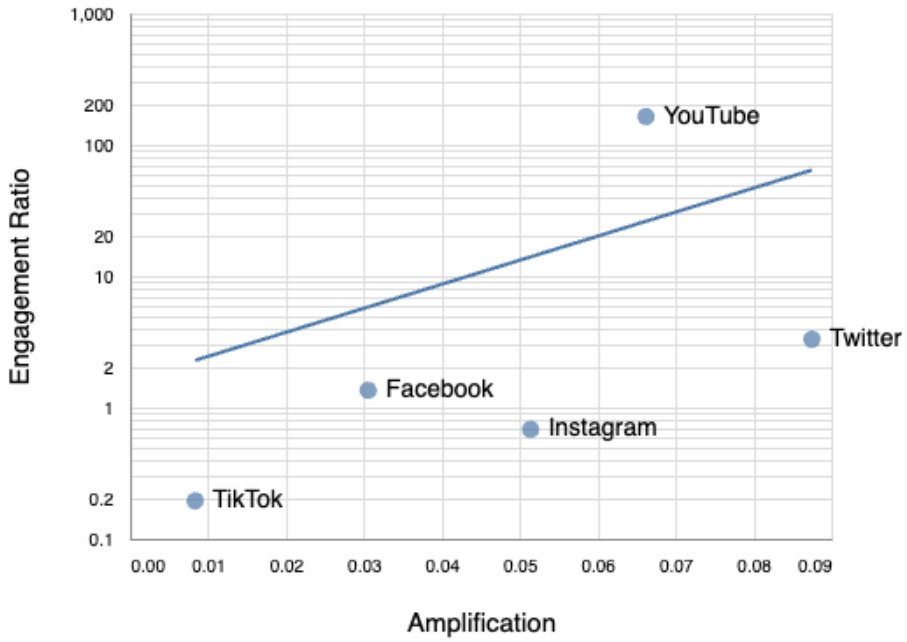

Figure A.6.11.14

This chart shows the positive correlation of the engagement and amplification. To normalise the engagement, an engagement ratio was used which is the median engagement (sum of the likes, shares, comments, and followers) of TVE content divided by the median engagement of non-TVE content. Amplification is the average proportion of the feed that contains TVE.

(note the log scale for the engagement ratio)

6.11.15 Removal Rate and Amplification

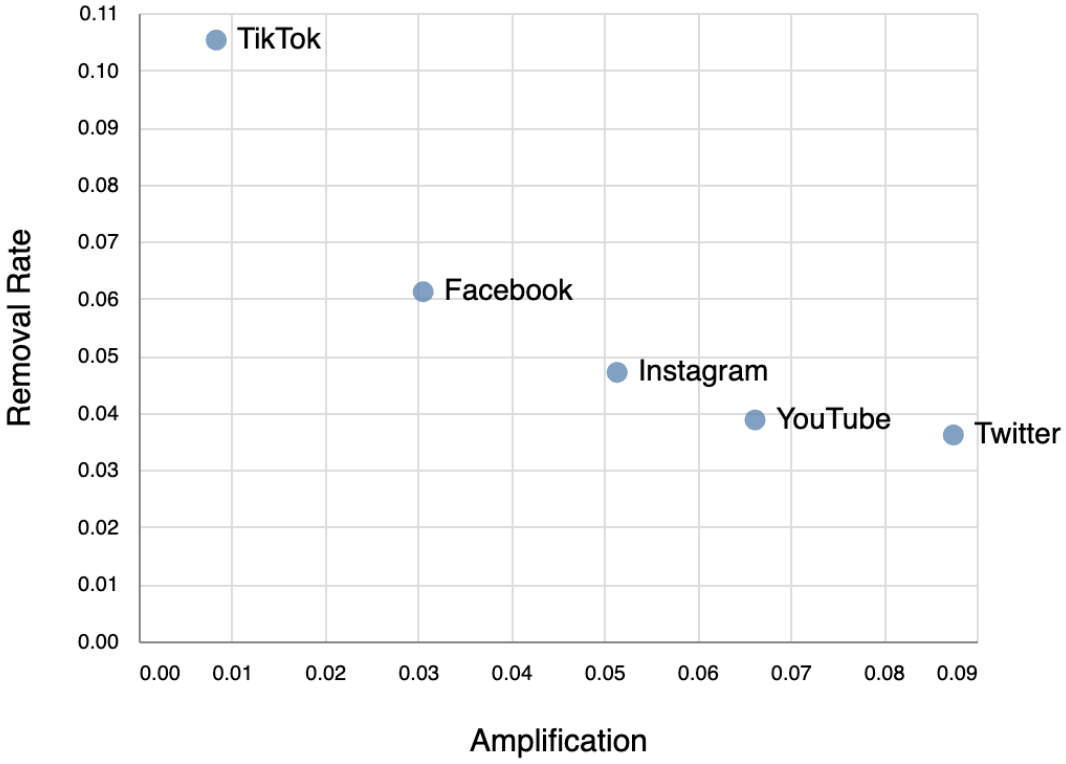

Figure A.6.11.15

This chart shows the inverse correlation of the removal rate and amplification.

6.12 Borderline Charts

6.12.1 Removal Rates per Platform

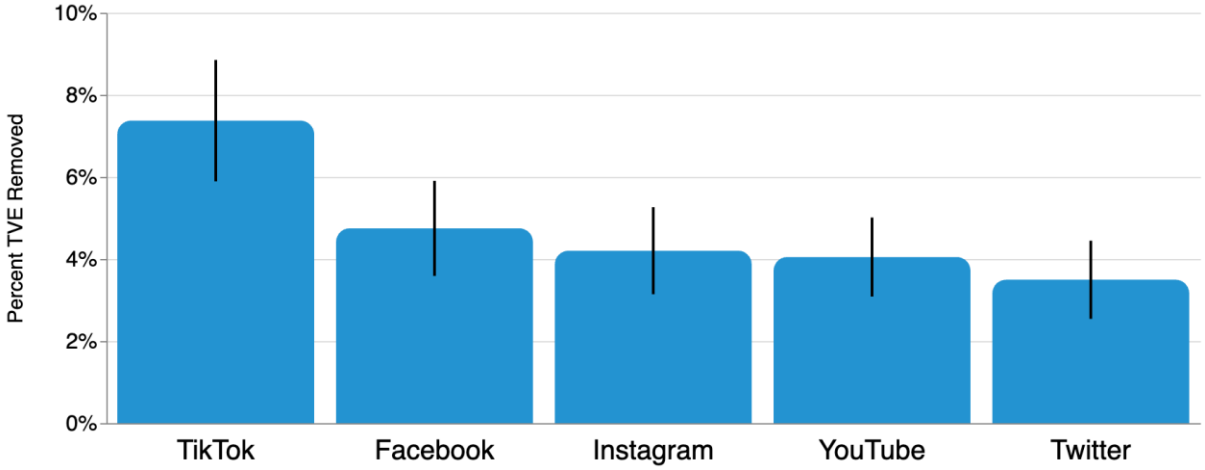

Figure A.6.12.1

This chart shows the percentage of Borderline TVE content that was removed by each platform. The black lines in the bars represent 90% confidence intervals.

	Facebook	Instagram	TikTok	Twitter
YouTube	0.6397	0.9144	0.0505	0.6846
Twitter	0.4009	0.6201	0.0237	
TikTok	0.1604	0.0774		
Instagram	0.7283			

Table A.6.12.1

P-values on Removal Rates per Platform (Borderline)

alpha = 0.01 (Bonferroni correction from 0.05)

6.12.2 Removal Rate per Language

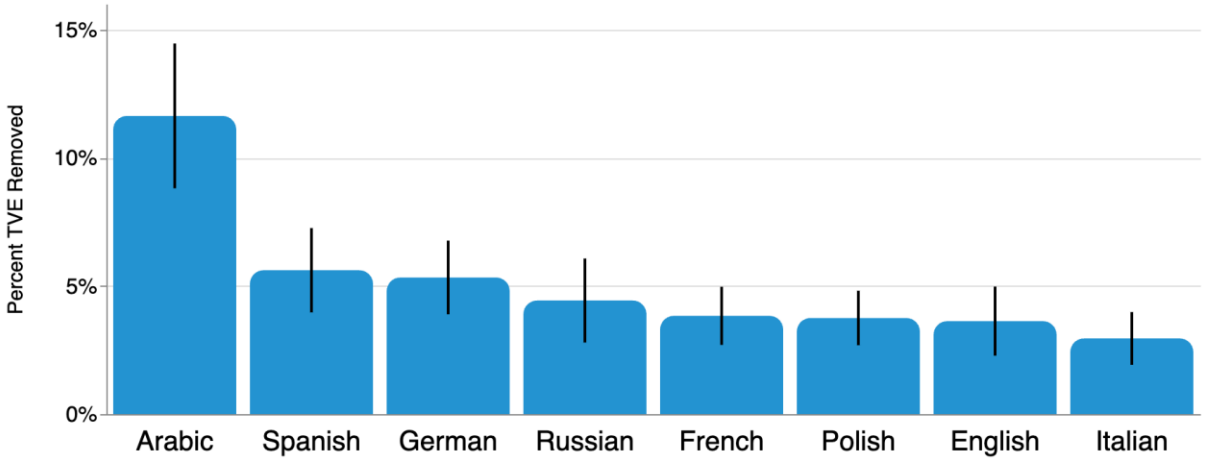

Figure A.6.12.2

This chart shows the percentage of Borderline TVE content that was removed for each language. The black lines in the bars represent 90% confidence intervals.

	Arabic	English	French	German	Italian	Polish	Russian
Spanish	0.0507	0.3527	0.3581	0.8964	0.1516	0.3207	0.6160
Russian	0.0226	0.7029	0.7601	0.6869	0.4226	0.7209	
Polish	0.0014	0.9431	0.9576	0.3679	0.5937		
Italian	0.0005	0.6862	0.5672	0.1741			
German	0.0281	0.3993	0.4095				
French	0.0023	0.9076					
English	0.0052						

Table A.6.12.2

P-values on Removal Rates per Language (Borderline)

a = 0.00625 (Bonferroni correction from 0.05)

6.12.3 Removal Rate per TVE Type

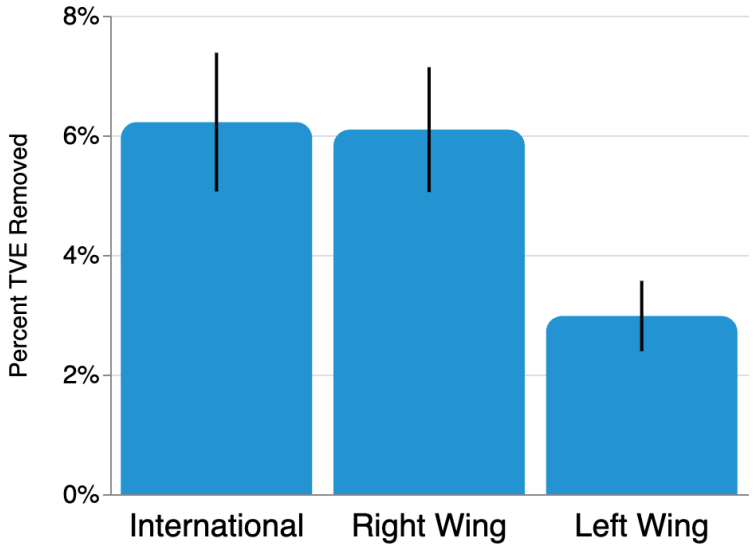

Figure A.6.12.3

This chart shows the percentage of Borderline TVE content that was removed for each TVE Type. The black lines in the bars represent 90% confidence intervals.

	International	Left Wing
Right Wing	0.9357	0.0051
Left Wing	0.0056	

Table A.6.12.3

P-values on Removal Rates per TVE Type (Borderline)

alpha = 0.017 (Bonferroni correction from 0.05)

6.12.4 Removal Time per Platform

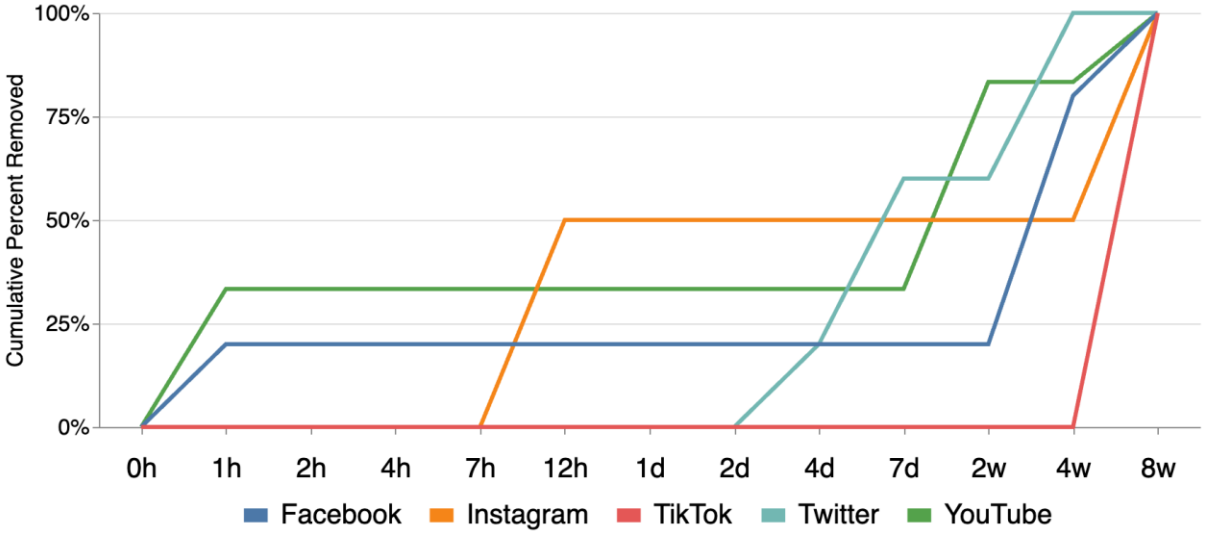

Figure A.6.12.4

This chart shows how long it took for each platform to remove the total amount of Borderline TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any Borderline TVE content that wasn't removed.

6.12.5 Removal Time per Language

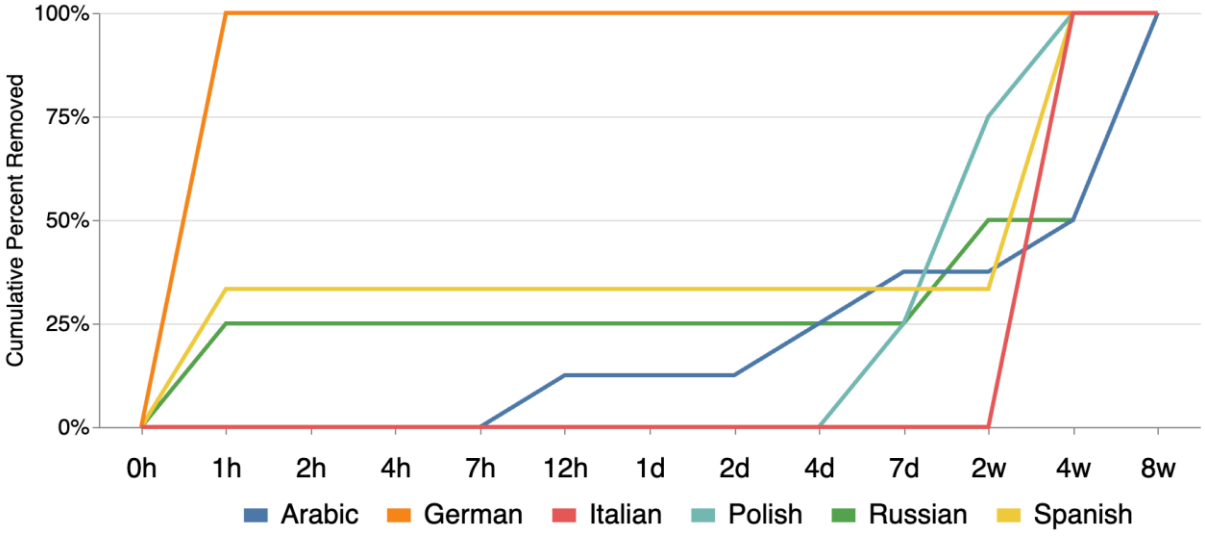

Figure A.6.12.5

This chart shows how long it took for each language to remove the total amount of Borderline TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any Borderline TVE content that wasn't removed.

6.12.6 Removal Time per TVE Type

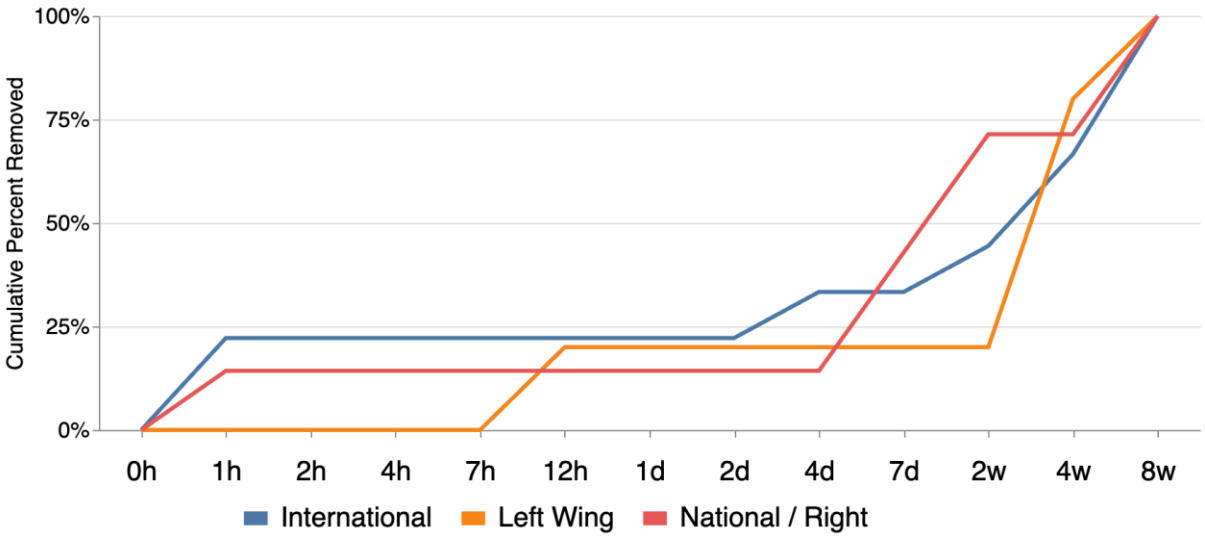

Figure A.6.12.6

This chart shows how long it took for each TVE Type to remove the total amount of Borderline TVE content. The line graphs show the cumulative percentage, ending in 100% at the top right corner. This chart does not take into account any Borderline TVE content that wasn't removed.

6.13 Task 3: List of References

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. New Media & Society, 18(3), 410-428.

Eirinaki, M., Gao, J., Varlamis, I., & Tserpes, K. (2018). Recommender systems for large-scale social networks: A review of challenges and solutions. Future Generation Computer Systems, 78, 413-418.

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. applied sciences, 10(21), 7748.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1), 2053951719897945.

Grimmelmann, J. (2015) The virtues of moderation. Yale Journal of Law & Technology 17: 42.

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian informatics journal, 16(3), 261-273.

Mohamed, M. H., Khafagy, M. H., & Ibrahim, M. H. (February 2019). Recommender systems challenges and solutions survey. In 2019 international conference on innovative trends in computer engineering (ITCE) (pp. 149-155). IEEE.

Murthy, D. (2021). Evaluating platform accountability: terrorist content on YouTube. American behavioral scientist, 65(6), 800-824.

Suhaim, A. B., & Berri, J. (2021). Context-aware recommender systems for social networks: review, challenges and opportunities. IEEE Access, 9, 57440-57463.

Zhang, Q., Lu, J., & Jin, Y. (2021). Artificial intelligence in recommender systems. Complex & Intelligent Systems, 7, 439-457.

END OF THE DOCUMENT

Publications Office
of the European Union