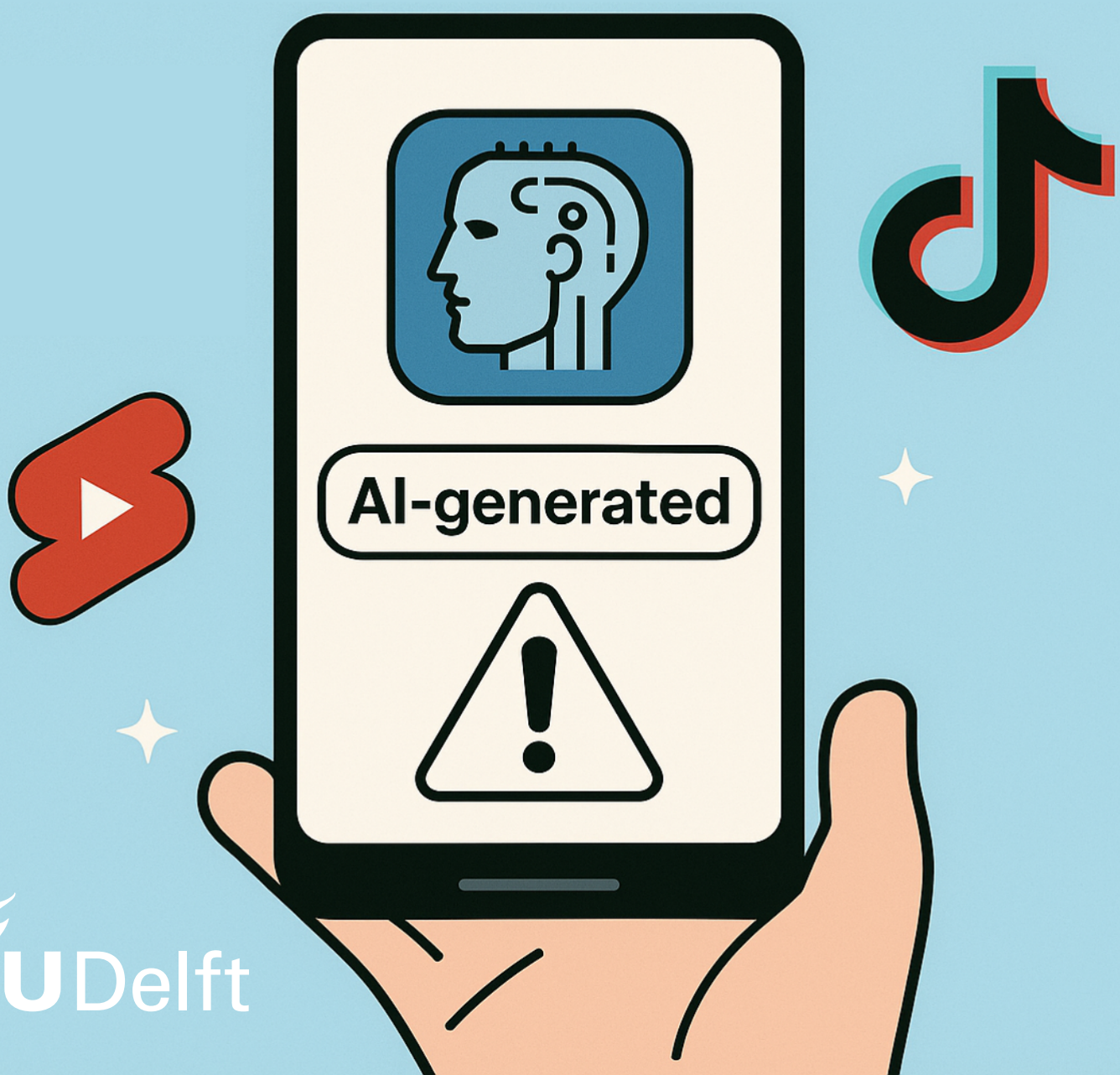
Labeling AI-generated Content on Short-Form Video Platforms

A Cross-Platform Analysis of YouTube Shorts and TikTok

EPA2942: Master Thesis EPA

Maaike Kuipers



Delft University of Technology

TU Delft

Labeling AI-generated Content on Short-Form Video Platforms

A Cross-Platform Analysis of YouTube Shorts and TikTok

by

Maaike Kuipers

4882466

Chair:	Prof. dr. M.E. (Martijn) Warnier (TU Delft)
First Supervisor:	Dr. S. (Savvas) Zannettou (TU Delft)
External Supervisor:	Rogier Schröder (KPMG)
Project Duration:	February, 2025 - June, 2025
Faculty:	Faculty of Technology, Policy and Management
Program:	MSc Engineering and Policy Analysis

Cover: Illustration generated with OpenAI DALL·E (ChatGPT, 2025)
Style: TU Delft Report Style, with modifications by Daan Zwaneveld

TU Delft

Executive Summary

In recent years, the rise of generative AI has fundamentally changed the way digital content is created and consumed. Whereas it used to be relatively easy to recognize whether content was created by a human, today's advanced AI tools have blurred this distinction. In response, leading short-form video platforms, including TikTok and YouTube Shorts, have implemented AI labels to indicate when content has been created or altered using artificial intelligence. However, the extent to which these labels are consistently applied, their visibility to users, and their influence on user engagement remain unclear.

This thesis investigates how AI labeling practices differ between YouTube Shorts and TikTok, and exploders how the presence of such labels relates to user behavior. A dataset of over 13,000 videos—7,092 from TikTok and 6,071 from YouTube Shorts—was collected using 25 AI-related hashtags. Quantitative analyses were combined with qualitative assessments of platform guidelines and labeling practices. The Mann-Whitney U test was conducted to assess whether engagement metrics significantly differed between labeled and non-labeled videos.

The results reveal notable platform differences in both prevalence and visibility of AI labels. On YouTube Shorts, labels appeared on 51.9% of videos but were almost exclusively placed in the expanded description (only 0.41% appeared in the player). For TikTok, 36.9% of videos included a label. These labels were displayed prominently beneath the creator's username. Most TikTok labels were creator-applied; platform-applied labels were rare.

Assessment of label presence against platform-specific guidelines showed contrasting outcomes on the two platforms. On YouTube Shorts, only 31.8% of videos carrying an AI label actually met YouTube's own criteria (precision), and just 56.6% of all videos had the correct label status overall (accuracy). In contrast, on TikTok labels—precision reached 98.1%, and 75.0% of videos were correctly labeled or left unlabeled according to TikTok's rules (accuracy). Nonetheless, almost one quarter of TikTok videos that met the platform's labeling criteria remained without any label.

Analysis of engagement metrics revealed small but statistically significant median differences across like/view ratio, comment/view ratio and views on both platforms. Videos without AI labels had higher median like/view ratios. YouTube Shorts showed lower median comment/view ratios for labeled videos while TikTok showed higher medians. Although these differences were statistically significant, their effect sizes were negligible for all metrics except one: only the median view count for labeled TikTok videos was significantly higher with a small, non-negligible effect size.

Overall, this research demonstrates that while AI labels have been introduced as a transparency measure, their practical implementation remains inconsistent, and their visibility to users is often limited. To address these challenges, it is recommended that platforms improve the visibility and clarity of AI labels and explicitly communicate whether a label is applied by the platform or the content creator. Standardizing labeling policies across platforms could further reduce confusion and lead to more consistent and accurate labeling practices. At the policy level, stronger regulations are needed to ensure that AI-generated content is not only detectable but also clearly disclosed to users. Finally, future research should investigate how evolving labeling technologies and stricter enforcement mechanisms may improve transparency and help users better navigate the growing presence of AI-generated content.

Contents

List of Figures

List of Tables

1

Introduction

In recent years, the rise of generative AI has fundamentally changed the way digital content is created and consumed. Where it was once easy to recognize whether content was created by a machine or a human, today's advanced AI tools have blurred this distinction. Recent studies confirm that users increasingly struggle to reliably identify AI-generated content [10, 22, 72]. Meanwhile, the growing competition among major tech companies is driving the rapid advancement of AI tools [19]. These ongoing developments are expected to further complicate the identification of AI-generated content in the near future.

Short-form video platforms play a central role in the spreading of this AI-generated content. In 2024, YouTube and TikTok reported over 2 billion and 1.5 billion monthly active users [21]. Their global reach amplifies both the societal influence of such content and the urgency of ensuring transparent communication about its origin.

In response to these rapid developments of Artificial Intelligence, the European Commission has adopted the EU AI Act [25]. However, this legalization primarily targets content generation tools and places limited responsibility on platforms which distribute this content. To enhance transparency, TikTok and YouTube Shorts have implemented AI labels that inform users when content has been modified or generated by AI. These labels are added manually by content creators or automatically by the platforms, based on machine-readable metadata required under the EU AI Act.

However, AI labels vary significantly across platforms in terms of visibility, wording, and source. This raises questions about their effectiveness and practical implementation. For example, TikTok uses the label "AI-generated content" and displays it in the video player. YouTube Shorts usually uses the label "modified or synthetic content," which is often displayed in less prominent places. While prior qualitative studies have explored user perceptions of AI labeling [12, 97], there is limited research done on the actual frequency and consistency of AI label implementation across platforms.

Furthermore, the effect of AI labels on user behavior remains unclear. Some studies suggest that labeled content is perceived as less credible and receives lower engagement [2]. Other studies indicate that labels may stimulate curiosity and increase user interaction [24]. These mixed findings underline the importance of platform context and highlight the need for quantitative, comparative research in the area of AI labels on short-form video platforms.

This research investigates how AI labeling practices differ between YouTube Shorts and TikTok, both in terms of their prevalence and the consistency of their implementation. It also explores the relationship between label presence and user engagement. To address these issues, the study applies both quantitative methods (e.g., descriptive statistics, hypothesis testing) and qualitative techniques (e.g., manual label validation and thematic coding). By combining these analytical approaches to study a complex, multi-actor policy problem related to AI labeling, the research aims to contribute to informed policymaking and improved platform regulation. The involvement of the EU AI Act, platform governance, and the potential impact of the increasing realism of AI-generated content on the society make this topic highly relevant to the field of Engineering and Policy Analysis.

Therefore, this study addresses the following main research question:

How do AI labeling practices on TikTok and YouTube Shorts differ in prevalence and consistency, and how do these labels impact user engagement?[1]

To answer this main research question, four sub-questions have been formulated:

1. What content moderation techniques are currently applied on short-form video platforms, and what are their main challenges according to literature?
2. How are AI labels applied on TikTok and YouTube Shorts, what are their characteristics, and what does existing research say about their application and user perception?
3. How prevalent are AI labels on TikTok and YouTube Shorts for videos with selected hashtags, and how is their presence related to user engagement within each platform?
4. How accurately and consistently are AI labels applied in practice compared to platform-specific guidelines across and within short-form video platforms?

The outline of this thesis is as follows: Chapter 2 addresses Sub-Questions 1 and 2 through a structured literature review. This review explores existing research on content moderation strategies, the use of warning labels, and current practices and challenges related to AI labeling. Chapter 3 describes the data collection process, including the selection of relevant hashtags, the scraping of TikTok and YouTube Shorts data, and the preparation of the dataset for analysis. Chapter 4 outlines the analytical methods used to address Sub-Questions 3 and 4. This includes statistical analyses to examine their relationship with user engagement and thematic coding to evaluate their application consistency. Chapter 5 presents the outcomes of these analyses. Finally, Chapter 6 interprets the results in light of the existing literature and discusses their implications for platforms and policy development.

[1] Throughout this thesis, AI labeling/AI labels is used as a generic term covering both the implementation of TikTok's 'AI-generated content' label [90] and YouTube Shorts' 'altered or synthetic content' label [103].

2

Literature Review

This chapter presents the approach and findings of the structured literature review conducted to answer Sub-Question 1 and Sub-Question 2. This review mainly focused on content moderation, AI labeling practices, and related challenges on short-form video platforms. Scopus was used as the primary academic database as it provides a broad range of peer-reviewed articles. In addition, backward snowballing was applied during the review process. This means that when a referenced article in the bibliography of a reviewed study was considered relevant, it was also included in the literature for this thesis. Besides academic sources, official documentation from the reviewed platforms was consulted to include platform-specific information. For statistical data, Statista was used as an additional non-academic source.

This chapter first discusses the literature search strategy (Section 2.1), followed by the review of literature for Sub-Question 1 in Section 2.2 and for Sub-Question 2 in Section 2.3. Finally, Section 2.4 presents the knowledge gaps identified through this study.

2.1. Literature Search Strategy

This section explains how the literature was selected and reviewed. A structured and transparent approach was followed to make the review process both comprehensive and reproducible. This structured literature approach is visualized in Figure 2.1.

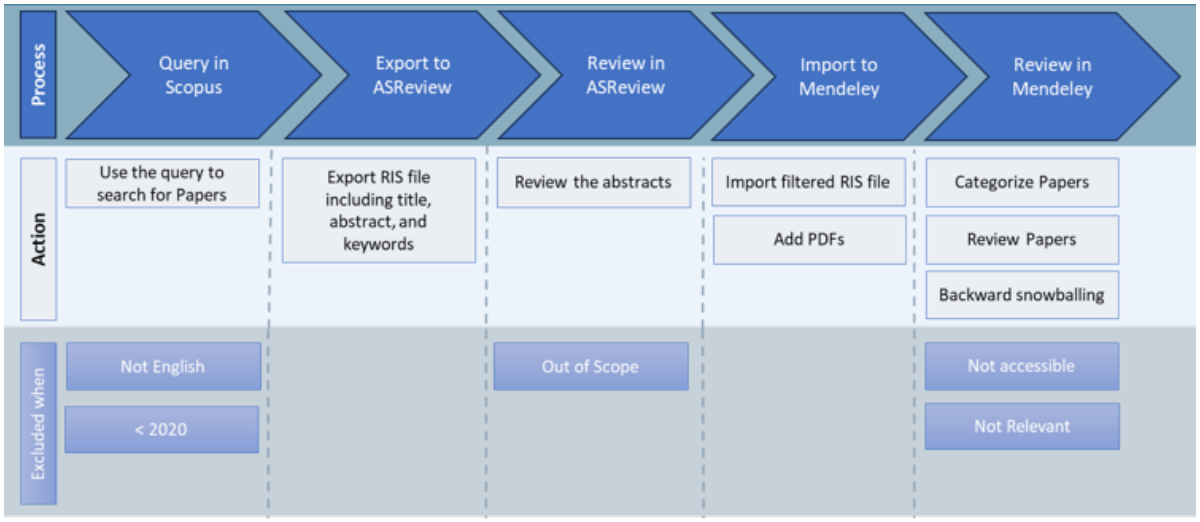


Figure 2.1: Structured Literature Review Approach

The same structured approach was applied to both sub-questions:

1. A search query was performed in Scopus (see Table A.1 for the specific queries).

2. Articles were excluded when:
    - The article had a publication year before 2020.
    - The article was not written in English.
3. The identified articles were exported as a RIS file from Scopus. This RIS file included citation information and abstracts.
4. The RIS file was imported into ASReview, an AI-powered tool designed to streamline article selection [4]. This tool allowed for the import of the RIS file with the abstracts. These abstracts were presented one by one for screening. Each abstract was classified as relevant or irrelevant to the sub-question. ASReview stored these decisions and made it possible to export a RIS file with only the articles which were classified as relevant.
5. This RIS file was then imported into Mendeley, a reference management software [59]. In Mendeley, articles were categorized into folders corresponding to each research question.
6. The full-text PDFs of the selected articles were added manually, followed by additional filtering based on the following criteria:
    - If an article was not publicly accessible, the TU Delft database was consulted.
    - If the article was unavailable in all sources, it was excluded.
7. Relevant parts of the studies were highlighted in Mendeley and added to a notebook in Mendeley where those quotes or parts were categorized.
8. Additional articles were identified through backward snowballing, where relevant studies found in the reference lists of included articles were included in the review.

2.2. Content Moderation and Its Challenges (Sub-Question 1)

This section reviews the literature to address the first sub-question:

What content moderation techniques are currently applied on short-form video platforms, and what are their main challenges according to literature?

Firstly, the concept and types of content moderation will be explained. Subsequently, important issues such as bias, transparency and inconsistencies in warning labels are discussed. Table A.1 presents the specific Scopus query and shows the number of articles identified, marked as relevant, and ultimately included.

The query included social media platforms and keywords associated with content moderation. The platforms included were the parent company of Instagram and Facebook (Meta), YouTube and TikTok. These platforms were selected because they are the most widely used short-form video platforms based on monthly active users [21].

2.2.1. Content Moderation

Combating disinformation on social media has long been a major challenge [32]. To address this issue, several regulatory frameworks have been introduced. Within the EU, the most prominent is the Digital Services Act (DSA). This framework includes stronger transparency rules for very large online platforms (VLOPs) [9, 26, 46].

In response to these regulatory pressures, content moderation strategies have been implemented by social media platforms. Content moderation refers to the process of monitoring or intervening in online discussions to ensure compliance with platform policies [99]. Moderation practices can be categorized based on their restrictiveness (hard vs. soft moderation), level of automation (automated vs. manual moderation), and governance structure (platform-based vs. community-based moderation). This section examines the current moderation techniques employed by platforms by highlighting both their effectiveness and the challenges they face.

Hard vs. Soft Moderation

One of the main differences in content moderation is between hard and soft moderation. Hard moderation refers to the outright removal of content or accounts. This is often in response to serious policy violations such as hate speech or misinformation. Examples of hard moderation are the de-platforming

of U.S. President Donald Trump on X and the removal of harmful content related to misinformation [40, 42]. Since these techniques typically occur after content has been seen by users, they are considered post-exposure interventions [23].

Soft moderation, in contrast, seeks to reduce the visibility or perceived credibility of content without removing it entirely. A widely used soft moderation strategy is the application of warning labels, which provide contextual information without restricting access [5]. These labels function as pre-exposure interventions, ensuring that users are aware of potential misinformation before engaging with content [23]. Warning labels have been applied across multiple domains, including COVID-19 misinformation [40, 51], (e-)cigarette health risks [49, 68, 81], abortion-related content [78], self-harm awareness [73], and political misinformation [62].

Manual vs. Automated Approaches to Moderation

Manual moderation involves human moderators who review flagged content. However, research shows that these moderators often work under poor conditions and face psychological stress due to prolonged exposure to disturbing material [33, 95]. To reduce this burden, many platforms have adopted automated moderation systems. These AI-based systems can automatically detect and flag harmful content. Yet, their effectiveness is limited, particularly when it comes to understanding context, sarcasm, and implicit bias [36, 75].

Because of these limitations, AI moderation is not accurate enough to operate independently, especially in detecting nuanced violations such as coded hate speech or misleading narratives [95]. As a result, most platforms use a semi-automated approach that combines automated content filtering with human oversight [43]. This hybrid model remains the dominant strategy, as it aims to balance the speed and scale of automation with the contextual understanding of human moderators.

Platform-Based vs. Community-Based Moderation

Another key distinction in content moderation practices is between centralized (platform-controlled) moderation and decentralized (community-based approaches). Platform-based moderation refers to moderation policies and enforcement mechanisms implemented directly by social media companies.

In contrast, community-based moderation is user-driven. This method relies on peer enforcement through reporting systems, comment flagging, and volunteer moderators. As Seering [77] argues, incorporating community-based moderation is important because it empowers users to self-regulate their online spaces. Nowadays, most large platforms rely on a combination of both community-based and centralized moderation strategies [38, 91]. However, their enforcement remains largely centralized, with final decisions made by platform authorities rather than community members.

Remarkably, Meta recently fired all third-party moderators (often referred to as fact-checkers) due to concerns about political bias. The company has since adopted a similar system to X, which employs a decentralized community note system [58, 94].

2.2.2. Criticisms and Challenges in Content Moderation

Although the previous section described the main types of content moderation strategies, their practical implementation has not been without problems. In practice, these strategies are often criticized for being inconsistent, opaque or even discriminatory. This section examines the most prominent challenges short video platforms face in implementing moderation practices, including bias and exclusion, user resistance, inconsistent labeling and lack of transparency.

Bias, Exclusion and Limitations of Content Moderation

A lot of major platforms shifted to the strategy of 'reducing' rather than 'removing' over the last years. Meta and YouTube applied this by reducing the visibility of flagged content [44, 74]. However, these platforms seem to struggle with ensuring fairness and transparency in their moderation processes.

Research has shown that TikTok's moderation system disproportionately impacts marginalized communities, particularly LGBTQ+ and black creators. These users report that their content is being deprioritized for no apparent reason, which is also known as shadow banning [39, 83]. Similar concerns have been raised regarding YouTube and Meta, which are also mentioned in the literature for employing comparable shadow banning practices [20, 44, 45]. Moreover, algorithmic moderation does not perform equally well across different languages and cultural contexts. This leads to weaker enforcement in non-English-speaking regions, further exacerbating existing inequalities on global platforms [63].

In addition to social and cultural bias, political bias has also been observed in moderation practices. Studies indicate that right-wing content is more frequently removed. This pattern is often attributed to the presence of hate speech, although this remains a highly debated and politically sensitive issue [43].

Finally, current moderation strategies have also shown limitations in terms of technical accuracy. For example, research found that Instagram's classifier was able to detect only 27% of the comments that users perceived as toxic [47].

In response to shadow banning and perceived bias in moderation systems, users have developed strategies to avoid detection. One example is the use of "algospeak" on TikTok. Algospeak refers to intentionally changing words (e.g., "le$bean" instead of "lesbian") to avoid being flagged or hidden [45].

In addition to changing their language, some users choose to migrate to alternative platforms with looser moderation policies. Inaccurate information removed from platforms like TikTok or YouTube often resurfaces on less regulated sites like BitChute and Odysee. On these platforms, engagement with such content remains high [64].

Labeling inconsistencies and Transparency Issues

A study by Ling, Gummadi, and Zannettou [51] on TikTok's COVID-19 warning labels found several inconsistencies. In some cases, unrelated content was incorrectly labeled, while harmful posts were not labeled at all. These errors can confuse users and reduce the effectiveness of labeling strategies.

A related problem is the so-called "implied truth effect." This means that users often see unlabeled content as more trustworthy, simply because they assume that if there is no warning the content must be accurate [32, 67]. On the other hand, wrongly labeling legitimate content can also damage trust, especially in credible sources. In addition, Zhang et al. [106] warns that flawed moderation systems may do more harm than good.

These labeling problems are part of a broader concern about transparency on social media platforms. According to their own statements, companies like TikTok and YouTube comply with the EU Digital Services Act (DSA). However, it often remains unclear how these platforms apply moderation and labeling in practice, especially when algorithms are involved [46, 74].

Even when platforms attempt to adjust their algorithms to promote reliable sources, research shows that institutional actors such as the WHO or NHS are frequently overshadowed by popular independent creators. This suggests that visibility on these platforms continues to be driven more by popularity than by expertise [54].

Meta has done some efforts to enhance transparency, such as the creation of its oversight board. However, this has also faced criticism. Researchers argue that these efforts often lack real accountability and diversity [8, 56, 69, 98].

2.2.3. Conclusion

This section described the main content moderation strategies currently used on short video platforms. These strategies include both hard moderation (content removal and de-platforming), as well as soft moderation (warning labels that aim to inform rather than restrict access) [5, 42].

To apply these strategies, platforms make use of semi-automated systems combining AI detection with human oversight [43]. However, these implementations are found to have bias, and some groups feel like they are deprioritized without a reason. So, users often find ways to bypass moderation. One way to do this is through "algospeak" or by switching to less-regulated platforms [45, 64].

Transparency remains a major concern, as platforms frequently fail to disclose the criteria behind their moderation decisions [18, 46]. The EU Digital Services Act (DSA) has been established to improve accountability and transparency in content moderation. However, its practical impact remains limited, particularly when algorithmic decision-making is involved [26, 46].

In addition, research finds that warning labels are applied inconsistently [51]. This inconsistency may reinforce the implied truth effect, leading to higher perceived trustworthiness of unlabeled content. Incorrect labeling, in turn, can undermine trust in credible sources [32, 67].

In conclusion, this section answered the first sub-question by outlining current moderation strategies on short-form video platforms and identifying their main challenges. The findings point to an urgent need for more transparent, fair and consistent moderation practices to effectively combat misinformation and maintain user trust.

2.3. AI Labels on TikTok and YouTube Shorts (Sub-Question 2)

This chapter addresses Sub-Question 2 by examining how AI labels are implemented on short-form video platforms:

How are AI labels applied on TikTok and YouTube Shorts, what are their characteristics, and what does existing research say about their application and user perception?

It focuses on the technical and regulatory frameworks behind these labels and explores differences in their application, visibility, and impact. The technical implementation is specifically examined for YouTube and TikTok, as these are the platforms included in the empirical analysis of this research. Further justification for this platform selection is provided at the beginning of Chapter 3.

The literature review was based on a structured search using keywords related to three main categories: (1) AI-generated content and labeling on short-form video platforms, (2) the technique both TikTok and YouTube use to support AI labeling (C2PA) [92, 101], and (3) the potential influence of AI labels on engagement, prevalence, perception, and trust. The search query included terms such as "AI-generated," "synthetic media," "C2PA," "label," "engagement," and "trust," and targeted platforms like TikTok and YouTube. The full search query is shown in Table A.1, along with the number of articles identified, marked as relevant, and used.

2.3.1. AI Labels

AI labels are a type of warning label designed to alert users that certain posts have been generated, altered, or enhanced using artificial intelligence. They represent a form of soft content moderation. This approach does not remove or block content but adds contextual information to influence user interpretation, as discussed in Section 2.2. Their main goal is to increase transparency and reduce the likelihood that users will misinterpret AI content as human-created [52].

The reduction of misinterpretations is especially important given the growing difficulty of distinguishing AI-generated content from that created by humans in the current media landscape [14, 29]. Such challenges are reflected in research showing that users often fail to detect AI-generated media, whether it involves tweets [22], visual profiles [72], or even faces in deepfake videos [10]. These trends underline the growing importance of clear and consistent labeling. The need for such labels is further reinforced by the increasing ease and scale with which AI-generated content can now be produced [79]. This makes effective labeling especially important in sensitive domains like healthcare and politics, where the risks of misinformation are higher [61].

To address the growing concerns surrounding artificial intelligence, the European Union has introduced comprehensive legal frameworks. The Artificial Intelligence Act (AI Act) represents the first comprehensive legal framework worldwide focused specifically on artificial intelligence. The act requires AI-generating content tools to ensure that such content is clearly labeled in a machine-readable format [25, 37]. For example, this means that tools like ChatGPT or other AI content generators must include a label in the metadata of the generated content, allowing platforms and systems to automatically recognize and disclose its AI origin.

However, TikTok and YouTube do not explicitly cite regulatory compliance as the main reason for introducing AI labels. Instead, they emphasize goals such as transparency, responsible AI use, and supporting both content creators and users [90, 92, 101]. Although these motives are consistent with EU policy objectives, both platforms present their initiatives as voluntary.

2.3.2. Standardization Through C2PA

To support transparency goals and ensure consistent labeling practices, both TikTok and YouTube are taking a dual approach to moderating AI-generated content. On the one hand, the platforms allow creators to voluntarily label their content as AI-generated. On the other hand, the platforms apply AI labels themselves by using the Coalition for Content Provenance and Authenticity (C2PA) standard [92, 101].

This standard enables platforms to read the machine-readable labels that AI content generation tools must include, as discussed in the previous chapter. These embedded labels contain metadata that specify whether the content has been generated or modified [6, 15, 87].

Figure 2.2 illustrates TikTok's automated AI labeling process based on the C2PA standard. The user generates content using an AI tool and uploads it to TikTok. During the upload, TikTok detects

the machine-readable metadata embedded in the file and automatically applies the ‘Generated with AI’ label.

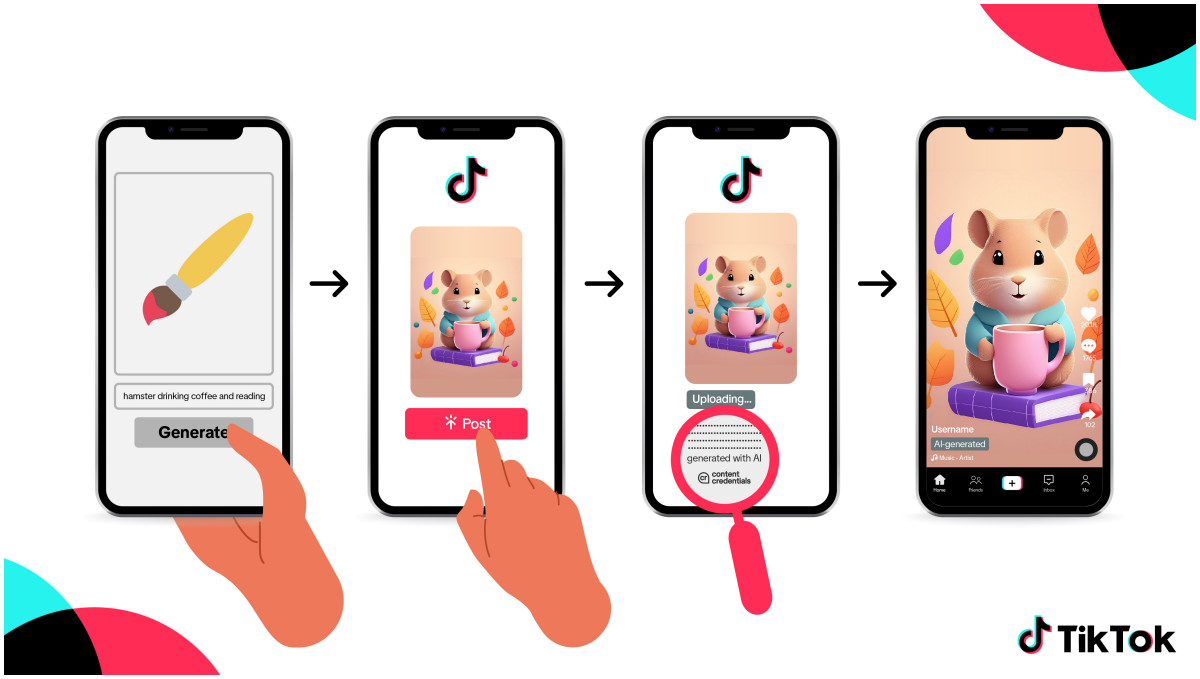


Figure 2.2: C2PA Label Application Process by TikTok. Source: [92]

2.3.3. Platform Differences in Labeling Practices

Although both platforms apply AI labels and rely on the C2PA standard, their approaches differ in terms of transparency of the source, placement, and disclosure requirements.

AI Labels on TikTok

TikTok applies two types of AI labels. The first is automatically added when AI-generated or manipulated content is detected via embedded C2PA metadata (see Figure 2.2). This ‘AI-generated’ label appears directly below the username (Figure 2.3a).

The second type is creator-applied and appears in the same place. If users voluntarily disclose the use of AI, TikTok adds the label ‘Creator-labeled as AI-generated’ (Figure 2.3b) [92].

AI Labels on YouTube Shorts

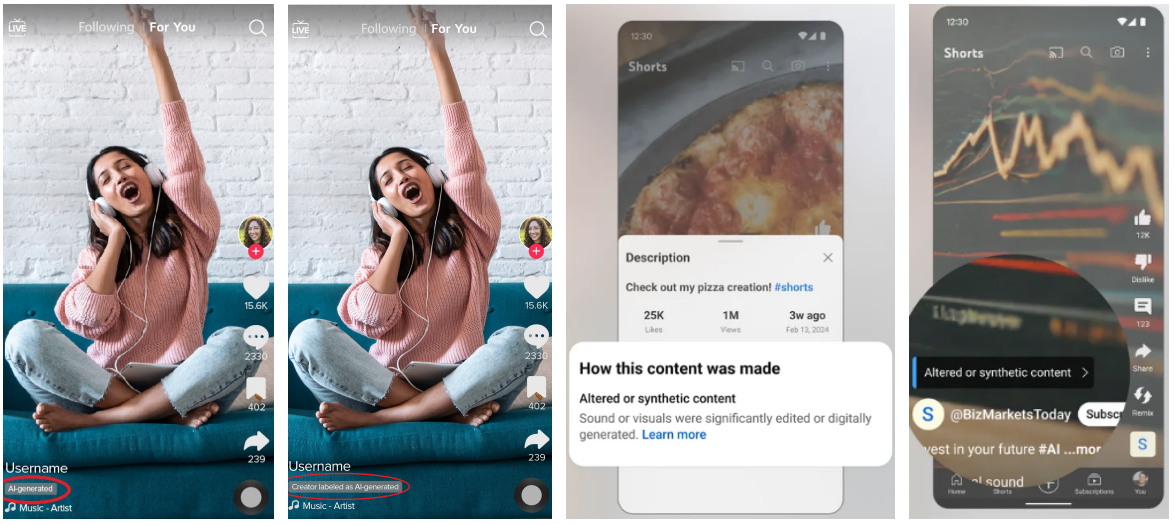
YouTube also applies AI labels, but its approach differs slightly. For example, it uses a different term: “Altered or Synthetic Content”. This label appears in the expanded video description (Figure 2.3c). For sensitive topics, the label also appears directly in the video player (Figure 2.3d) [101].

Creators can choose to label their content, but unlike TikTok, YouTube does not indicate whether a label was added by the creator or the platform. Additionally, not all AI-generated content requires labeling; clearly unrealistic content does not require a label [101].

Comparison of AI Labeling Practices

Although both TikTok and YouTube rely on the C2PA standard, they differ in how they implement and display AI labels.

- Transparency of Source: TikTok distinguishes between platform-applied and creator-applied labels, while YouTube does not.
- Label Text: TikTok uses AI-generated, while YouTube uses synthetic or altered content.
- Label Placement: TikTok’s labels appear in the main interface, directly under the username. YouTube shows the warning only in the expanded description unless the topic is sensitive.
- Disclosure Requirements: On YouTube, clearly unrealistic content does not require labeling. However, on TikTok, nearly all AI-generated content must be labeled.



(a) AI Label Applied by TikTok . Source: [90].

(b) AI Label Applied by TikTok Creator . Source: [90].

(c) AI Label in the Expanded Description on YouTube Shorts . Source: [101].

(d) AI Label Visible in the Video Player on YouTube Shorts (Sensitive Topic) . Source: [101].

Figure 2.3: Platform-applied AI labels on TikTok and YouTube

These differences reveal broader inconsistencies in how AI-generated content is labeled across platforms, despite shared technical foundations and regulatory pressures. While platform policies and technical frameworks determine how and if AI labels are applied, their actual effectiveness depends on how users perceive and interact with labeled content.

2.3.4. Engagement, Trust, and Transparency Effects of AI Labels

A growing body of research examines how AI labels influence user perception and engagement, but findings remain mixed. Some studies suggest that people perceive content without AI labels as more authentic [11, 65], while AI labeled content is often viewed as less credible [52] and shared less frequently [2]. Although this reduction in sharing may seem counterproductive for user engagement, it can help limit the spread of potentially misleading synthetic media [7].

However, Li and Yang [50] found no significant main effects of AI labels on perceived sharing intentions. Yet, they observed significant interaction effects, indicating that the effectiveness of such labels depends on the content type and informational context.

Adding to this complexity, some studies report positive effects of AI labeling. A research shows that labeling AI-generated content can significantly increase behavioral engagement even though it does not affect users' psychological perceptions Du, Zhang, and Ge [24]. This increased engagement is attributed to users' curiosity about AI-generated content, which motivates them to interact more frequently with labeled material.

Despite these insights, most quantitative research on AI labeling focuses on text-based content, and studies examining its effects on user behavior on short video platforms remain limited. Although some research has explored user engagement with AI-generated images, such as paintings and photographs [96], the role of labeling in these contexts has not been examined.

In addition to these quantitative studies, qualitative research has explored which types of AI labels are considered most appropriate and effective for different user groups [12, 97]. These studies underscore that it is important to design AI labels that are both clear and consistent, while also accounting for platform-specific characteristics, target audiences, and content types.

Beyond user engagement, trust and perceived transparency also play a key role in how effective AI labels are. Studies show that trust in such labels varies across user groups and contexts. For example, young adults are more likely to trust credibility markers issued by reputable organizations than those added by content creators themselves [80].

The ethical value of AI labeling further depends on how clearly and visibly this information is presented. Labels that are hard to notice or understand can create a false sense of transparency, allowing

deceptive practices to persist [31].

Finally, as generative AI becomes more advanced, concerns grow that it may reinforce social biases [76] and unintentionally shape public opinion [53]. In this context, clear and consistent labeling is an essential tool for content moderation, helping users better assess where content comes from and how it was created.

2.3.5. Conclusion

This section examined how TikTok and YouTube implement AI labels, highlighting differences in technical standards, label visibility, and enforcement criteria. Both platforms use the C2PA standard to support content traceability [15, 60]. However, their implementation differs significantly. TikTok applies both platform- and creator-assigned labels, which appear visibly in the main interface [90]. YouTube, by contrast, usually shows labels only in the video description or in the player for sensitive topics. It also does not indicate whether the label was applied by the platform or the creator [101].

Existing research on AI labels shows that user perception is highly context-dependent. Labeled posts are often perceived as less trustworthy [2, 52], and may be shared less frequently. However, some studies find the opposite effect: labeling can increase behavioral engagement—such as likes and shares—driven by user curiosity about AI-generated content [24]. Trust in AI labels also varies by source; users generally place more trust in labels from platforms or institutions than in those added by individual creators [80].

Despite these developments, inconsistencies remain in how labels are applied and understood. Research highlights the importance of consistent and clear labeling practices to avoid the illusion of transparency [31], yet current platform strategies vary in execution and effectiveness.

This section answered Sub-Question 2 by providing an overview of how AI labels are technically implemented on TikTok and YouTube, what their main characteristics are, and how users perceive and respond to them according to existing research. Building on these insights, the next section discusses the key gaps found in the literature and presents how this study contributes to addressing them.

2.4. Knowledge Gaps

AI labels are increasingly used to improve transparency and inform users about AI-generated content. However, their implementation across platforms remains inconsistent in terms of visibility, criteria, and enforcement [90, 101]. While TikTok and YouTube both rely on the C2PA standard, they differ significantly in label placement, user disclosure requirements, and whether they distinguish between platform- and creator-applied labels.

Although several qualitative studies have explored perceptions and preferences of AI labels [12, 97], little is known about the actual prevalence of these labels and whether they are applied consistently and in accordance with platform-specific guidelines. Existing research tends to focus on related areas, such as misinformation labels [51] or AI content on other platforms like Pixiv [96], but rarely on AI labels themselves, and not in a cross-platform comparison.

Given prior findings on inconsistencies in warning label application [51], similar issues may arise with AI labels. Yet to date, no study has systematically compared how frequently AI labels are used, whether they follow platform rules, and how these factors influence user engagement, which is the focus of this research.

Furthermore, some studies focused on how users perceive AI labels in terms of credibility, authenticity, and trust [11, 65, 80]. A smaller but growing number of studies has examined behavioral engagement, such as likes, shares, and comments. The findings are mixed. Some studies report that labeled content receives less engagement because users see it as less credible [2]. Others find the opposite: labeled content can attract more interaction, possibly due to curiosity [24]. These differences show that the impact of AI labels depends strongly on the context and platform, and they underline the need for more focused research on short-form video.

In addition, user trust in AI labels is known to be context-dependent and influenced by factors such as label visibility and source credibility [32, 80]. These inconsistencies risk undermining transparency efforts and the implied truth effect may unintentionally reinforce misinformation believes [67].

Addressing these gaps is essential to understanding how AI labels work in practice, how they affect user behavior and how consistently they are applied. This study quantitatively examines the prevalence of AI labels and their relationship with user engagement on TikTok and YouTube. In addition, a qualita-

tive analysis of platform guidelines assesses whether these labels are applied consistently and in line with stated policies. The findings aim to contribute to academic debates and provide policymakers and the social media platforms with recommendations for more standardized, transparent, and effective labeling practices.

3

Data Collection

This chapter discusses the method used to collect the data for this thesis. In Section 2.2, multiple short-form video platforms were considered. Specifically, the four most widely used platforms were examined: YouTube, Instagram/Facebook (Meta), and TikTok [21]. Due to time constraints, this study focuses on two platforms. This allows for a cross-platform analysis while keeping the data collection process manageable. Including a third platform would have required the development and execution of an additional data collection pipeline, which was deemed too time-consuming.

YouTube Shorts was selected because it provides a publicly accessible API YouTube [102]. TikTok was chosen as it is the only major platform that exclusively features short-form video content, making it particularly relevant for this research.

3.1. Data Collection for Sub-Question 3

To answer Sub-Question 3, there are two types of information needed:

1. The presence of an AI label
2. Engagement metrics of a post (e.g. likes, shares, comments, views)

This section describes the approach used to collect the dataset for Sub-Question 3. The data collection existed of five main steps: (1) hashtag selection, (2) URLs collection, (3) metadata collection, (4) data merging, (5) data filtering, and (6) iteration (see Figure 3.1). These steps were repeated until approximately 13,000 videos (from both TikTok and YouTube Shorts) were collected using 25 AI-related hashtags. For each video, the process captured both the presence of an AI label and key engagement metrics (views, likes, comments, shares). The repeated six steps are outlined first, followed by a general explanation of how these steps were implemented in Python[1]. All data was collected between March 21 and March 31, 2025.

[1]All corresponding code is available at `https://github.com/maaikekuipers/thesis`. The repository includes a detailed README file explaining the installation of required Python packages and how to execute the scripts.

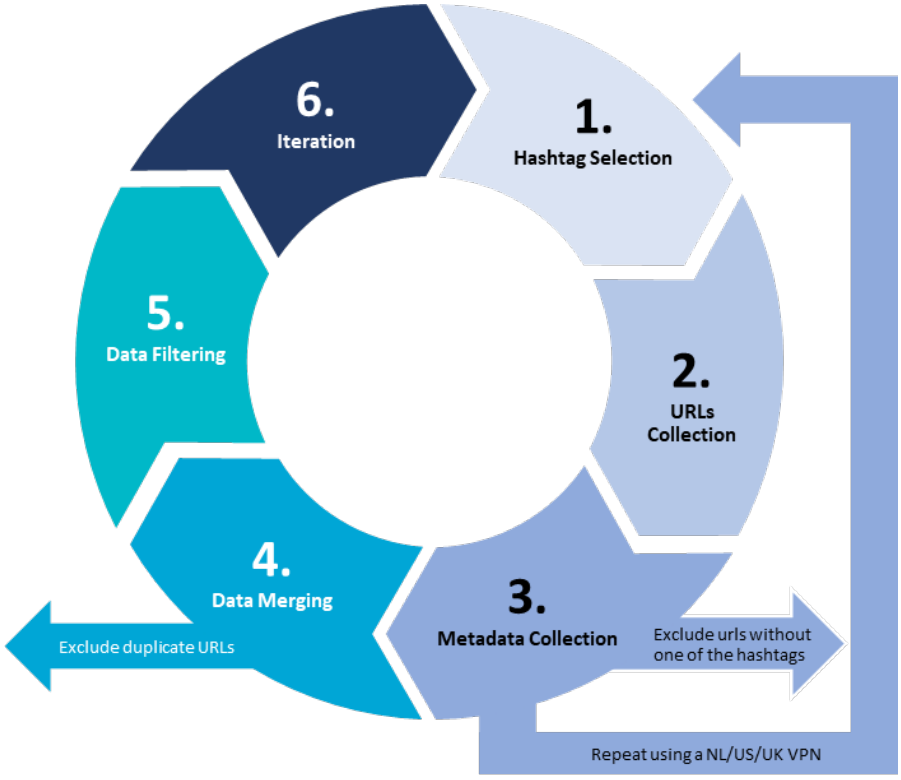


Figure 3.1: Overview of the Data Collection Cycle

3.1.1. Data Collection Cycle

Step 1: Hashtag Selection

Hashtags enable platforms to group related content together and are frequently used by creators to identify the theme of a video [41]. Since both YouTube Shorts and TikTok allow the usage of hashtags, these served as the primary entry point for collecting AI related videos in this study. This approach was also used in another research to collect COVID-19 related videos [51].

The data collection started with the hashtag `#ai`. This hashtag was chosen because it contained over 9 million videos on both platforms as of March 21, 2025 [88, 100]. In the following iterations, new hashtags were added based on their co-occurrence in the metadata of the retrieved videos. The 6 steps were repeated for 25 hashtags, resulting in a dataset of 13.163 YouTube Shorts and TikToks.

Step 2: URLs Collection

A custom scraping script was developed as part of this thesis to collect the data from both platforms. This script was implemented as uniformly as possible across platforms. It directly visits specific hashtag pages of TikTok or YouTube Shorts. Hashtag pages are publicly accessible web pages that display a feed of videos associated with a specific hashtag (e.g., `https://www.tiktok.com/tag/ai` or `https://www.youtube.com/hashtag/ai`). These pages were visited in private and logged out mode, to reduce the algorithmic bias in the data collection. When visiting these pages, it required manual CAPTCHA solving for TikTok. CAPTCHAs (Completely Automated Public Turing tests to tell Computers and Humans Apart) are mechanisms designed to distinguish between humans and automated systems [35]. For YouTube Shorts, manual cookie rejection was required. After manually completing the CAPTCHA (TikTok) or rejecting cookies (YouTube Shorts), the script scrolled through the hashtag page to collect video URLs. For this, the scrolling algorithm described by Steel and Abrahams [82] was used.

However, TikTok limits each hashtag page to approximately 250 videos per session. To bypass this limit and collect more data, a VPN-based approach was used. By simulating user locations in different countries, the same hashtag page could be accessed multiple times. Besides the home country (the Netherlands), VPN servers from the United Kingdom and the United States were used. These locations were chosen to maintain relevance to English-language hashtags. This method enabled the retrieval of

approximately 250 videos per hashtag from the Netherlands and the UK, and around 150 from the US. Although some overlap occurred between sessions, this approach substantially increased the overall volume of collected data.

To ensure robustness and prevent data loss, URLs were periodically saved to external files. This intermediate step served two primary purposes:

1. Saving URLs at regular intervals allowed to resume the metadata collection from saved files if crashes occurred.
2. Storing URLs enabled a comprehensive check with the finalized hashtag set.

These files were temporarily stored on the TU Delft project drive. Further elaboration on data storage and privacy considerations can be found in Section 6.3.1.

Scraping and API Usage
TikTok's official API is not publicly available and, despite a request, access was not granted. As a result, this study utilized an unofficial TikTok API [86]. For YouTube, the official API is publicly available [102].

Although both (un)official APIs support hashtag-based queries, these often returned large numbers of videos only loosely related to the specified hashtags. Since the aim was to include only videos explicitly tagged with at least one selected hashtag, this resulted in many API calls with limited usable results. A clear example of these limitations is YouTube's API. It imposes a strict daily quota of 10,000 calls, which was often exceeded using this method. In addition, the API frequently returned regular YouTube videos instead of Shorts. The only available query parameter to limit this was duration = 'short', which proved insufficient.

Given the scope of this thesis and the objective to compare YouTube Shorts with TikTok, this still resulted in many unnecessary API calls. This is the reason why there was a scraping script created instead of utilizing the platforms' APIs.

Step 3: Metadata Retrieval
For all collected URLs, metadata was retrieved using the official YouTube API and an unofficial TikTok API. The following metadata was collected:

- URL
- Views
- Likes
- Comments
- Publish Date
- Hashtags (extracted from the description and/or title)

For TikTok, shares and the AI label type were also extracted. The YouTube API did not provide this information. Therefore, a separate script was developed which checked all the URLs after completing the data gathering cycle.

URLs that did not contain any of the relevant hashtags were excluded from further processing. Steps 1–3 were repeated separately for each VPN location: the Netherlands, the United States, and the United Kingdom.

Step 4: Data Merging
The data from the three VPN locations were merged in a central Jupyter notebook. In this notebook, the number of videos per VPN location and the total size of the dataset was recorded. During the merging process, duplicate videos were automatically removed. Due to the size of the individual datasets and the final merged dataset, it became clear that a significant number of duplicates existed between countries.

Step 5: Data Filtering

In the same notebook, hashtags were extracted from the merged dataset. Variations in capitalization (e.g., #AI, #ai, #Ai) were standardized. Initially, the top 15 most frequently co-occurring hashtags were considered for expanding the hashtag set. However, during early iterations, this threshold sometimes resulted in no new relevant hashtags being identified. To address this and maintain growth of the dynamic hashtag set, the threshold was increased to the top 18 co-occurring hashtags. This adjustment successfully captured additional relevant AI-related hashtags that had not yet been included.

Only hashtags explicitly related to AI were retained. In cases where multiple variations of a hashtag existed (e.g., #AIvideo vs. #AIvideos), only the most frequently used variant was added to avoid duplication.

All identified hashtags and their classifications are presented in Section B.3. In these tables, the specific search hashtag is shown alongside its top 18 co-occurring hashtags. A color-coding was applied to visualize the filtering decisions:

- Green: Newly identified and relevant AI-related hashtags, which were immediately added to the dynamic hashtag set during the current iteration.
- Orange: Hashtags already identified through co-occurrence with a different search hashtag; these were not added again.
- Light Green: Hashtags similar in meaning or form to previously identified hashtags. These were not added.

Step 6: Iteration

The process was repeated for each newly identified hashtag until a total of 25 unique hashtags were collected. These hashtags are listed in Table 3.1.

Because TikTok returned a relatively small number of videos per country, each hashtag page was visited twice for every VPN location (Netherlands, United Kingdom, and United States). This approach enabled to expand the dataset.

In total, 13,163 videos were collected using the 25 selected hashtags. This included 7,092 TikTok videos and 6,071 YouTube Shorts.

Table 3.1: Final Hashtag Set Used for Data Collection

#ai	#aiart	#aigenerated	#aivideo	#artificialintelligence
#aicat	#aiedits	#aivfx	#aitools	#aibaby
#aifashionmodel	#aigeneratedfilm	#aistorytelling	#aistory	#creativeai
#aicontent	#aiinnovation	#aianimation	#aiproductions	#aitrends
#airevolution	#aicommunity	#aiartcommunity	#aiartwork	#aiartist

Final Hashtag Check

To ensure the completeness of the final dataset, an additional script was developed which identifies any video URLs that had been scraped earlier but were missing from the consolidated dataset. The script scanned all the the intermediate saved files (organized by hashtag and country) and compared them to the merged dataset.

This additional check was necessary because the hashtag set used to filter relevant videos was expanded iteratively during the data collection process. As a result, some videos that did not contain any of the selected hashtags at an earlier stage may have become relevant later, after new co-occurring hashtags were added. By rerunning these URLs through the updated hashtag set, previously excluded but now relevant videos were recovered.

3.1.2. Implementation of Data Collection scripts

Figure 3.2 displays a high-level overview of the scripts used in the data-gathering process. All scripts were written in Python and developed specifically for this thesis. They make use of the Playwright library and build upon the scrolling behavior and data extraction techniques described in the unofficial TikTok API documentation [86].The approach was adapted for use on YouTube Shorts as well.

Steps 1, 2, and 3 of the data collection cycle were implemented using a platform-specific `Hashtag Search` script. This script was configured manually by specifying both the target hashtag and the VPN

country, enabling automated navigation to the corresponding hashtag page and ensuring that the retrieved data was stored during the cycle. Depending on the platform, the script then called either the `YouTube API` or `TikTok API` module.

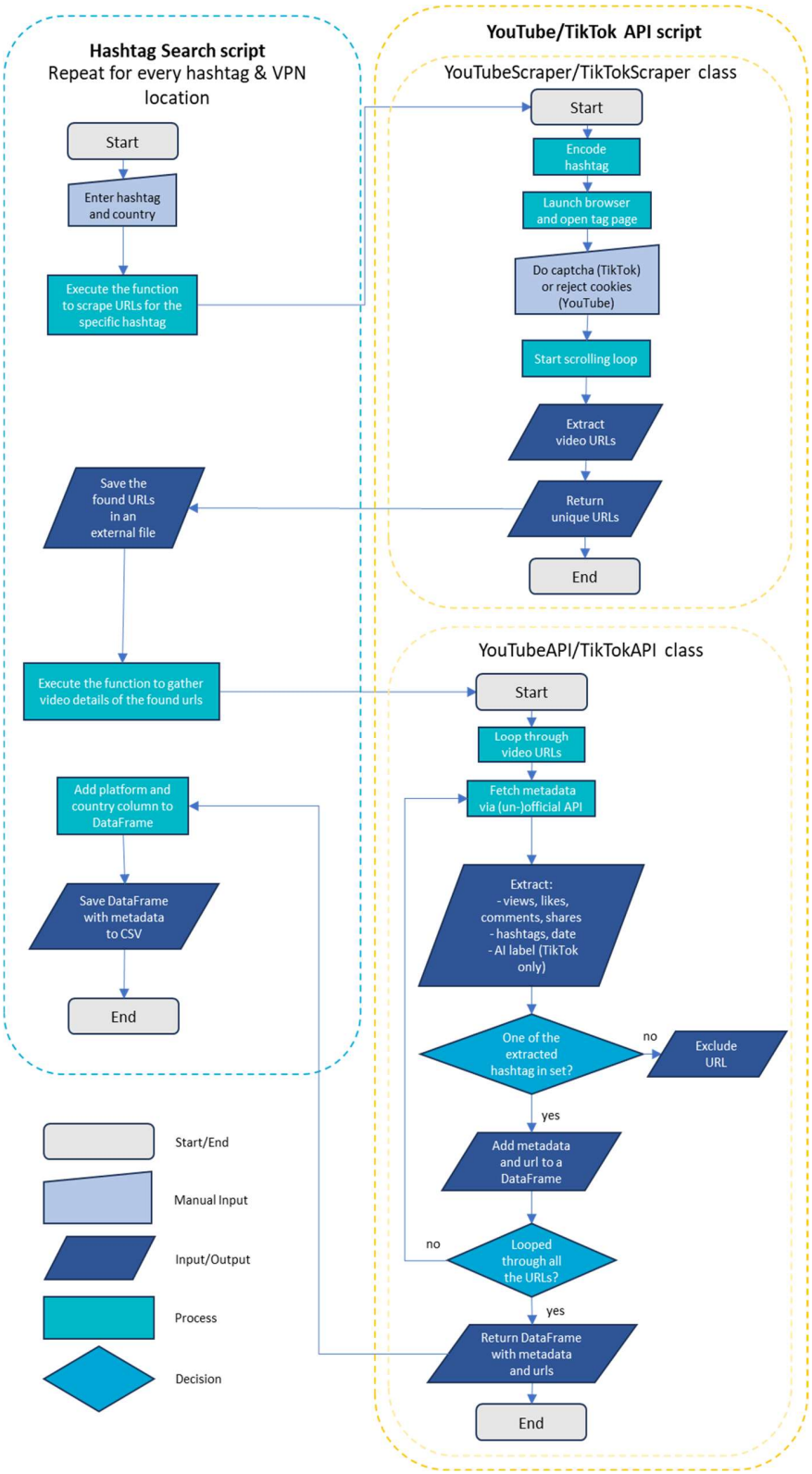


Figure 3.2: High-Level Overview of the Data Collection Process Scripts

These designed YouTube/TikTok API script consisted of two main components:

1. **Scraper class** (`TikTokScraper` / `YouTubeScraper`)
    - Encodes the hashtag, so it could be used in the hashtag page url (e.g., `https://www.tiktok.com/tag/ai` or `https://www.youtube.com/hashtag/ai`).
    - Utilizes Playwright to launch a browser and scroll through the platform-specific hashtag page and extract video URLs.
    - Scrolling continues until a target number of videos is reached, with delays added to avoid aggressive behavior. The scrolling is based on the method developed by Steel and Abrahams [82].
    - Returns the unique URLs to the main script, which executes the function to gather the metadata of the extracted URLs.
2. **API class** (`TikTokAPI` / `YouTubeAPI`)
    - Retrieves video metadata for the collected URLs via either the official YouTube API or the unofficial TikTok API.
        - Since the API responses contained large amounts of unstructured data, filtering was applied to extract only the relevant metrics.
    - Applies a filtering step to exclude videos that do not contain any of the selected hashtags.
    - Returns a DataFrame with relevant metadata such as views, likes, comments, publish date, and hashtags.
    - For TikTok, AI labels and share/save metrics are also included; YouTube Short AI labels are gathered later via a separate script.

After repeating this for all of the hashtags and the VPN locations, the information about the presence of AI Labels on YouTube were collected since that information was not provided by the YouTube API. For this, the third class within the `YouTube API class` was called (`LabelCheckerYouTube`). After inspecting the elements in the HTML code, there were two elements found which indicates whether an (sensitive) AI label was present. Therefore, this class looked for the following elements in the HTML code:

1. `ytwHowThisWasMadeSectionViewModelHost`
    - Indicates that an AI label has been applied in the expanded video description.
2. `<span class="ytwPlayerDisclosureViewModelText">Altered or synthetic content</span>`
    - Indicates that the content is related to a sensitive topic, such as politics or health, and has a visible label in the video player as well.

If either element was detected, the script recorded the presence of an AI label. The second element additionally triggered a separate flag indicating that the content related to a sensitive topic. Some videos had become inaccessible (e.g., deleted, set to private, or region-blocked) in the time between initial data collection and label retrieval, and were therefore excluded from further analysis. To avoid data loss, the dataset with and without YouTube labels were kept separate during this process and merged at once during the data cleaning phase.

3.1.3. Data Cleaning

Building on this prefiltered dataset, further data cleaning was conducted after the datasets had been merged. During the collection of AI labels for YouTube Shorts, the dataset without AI labels was kept separate from the dataset with added AI labels. This approach enabled a one-time merge, allowing for easier verification of the process and reducing the risk of errors during the update.

Merging Datasets

The datasets with and without YouTube AI labels were merged into one dataset. During this merging process, special care was taken to avoid accidentally deleting valid rows. The only rows that were explicitly deleted were those without AI labels. Further investigation revealed that these deleted rows corresponded to videos that had either been removed or made private on YouTube Shorts between

the initial scraping phase and the label-checking stage, as discussed in Section 3.1. Consequently, excluding these rows from the final merged dataset was considered justified. In addition, the merged dataset was checked for duplicate entries. None were found, indicating that the merging process was successful.

NaN likes

After validating some of the data points, several rows were found to have zero likes despite a high number of views in the non-labeled dataset. Upon further inspection, it was determined that YouTube had hidden the like counts (see Figure 3.3). Since the goal of Sub-Question 3 is to analyze the relationship between the dataset and engagement metrics, these rows were removed.

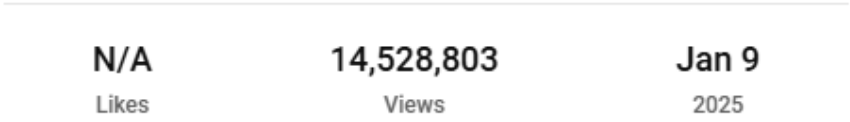


Figure 3.3: Na Likes on YouTube Shorts

Platform-Specific Metrics and Missing Values

To further validate the dataset, a check for missing values was performed. Two features showed systematic gaps due to platform-specific availability. First, the sensitive topic label was missing (NaN) for all TikTok videos, as this feature is not supported on that platform. In contrast, YouTube Shorts explicitly contained either a 0 (no label) or 1 (label applied). Second, the 'shares' engagement metric was missing from all YouTube Shorts, since this metric is exclusively provided by TikTok. Both features were retained to support within-platform analyses.

In addition, a manual check of a subset of the data revealed that TikTok distinguishes between two types of AI labels: a value of '1' for creator-applied labels and '2' for platform-applied labels. YouTube, by contrast, does not differentiate between these label types. To allow for meaningful comparisons between platforms while retaining TikTok's label detail, the labeling scheme was adjusted as follows:

- Creator-applied TikTok labels were recoded as '2', and platform-applied labels as '3'.
- For cross-platform comparisons, both label types can be aggregated and recoded as '1' to align with YouTube's undifferentiated labeling system.

The resulting label structure is summarized in Table 3.2.

Table 3.2: AI Label Structure in the Final Dataset

Label Status	YouTube	TikTok
No Label Applied	0	0
Label Applied by YouTube/Creator)	1	N/A
Label Applied by Creator	N/A	2
Label Applied by TikTok	N/A	3

VPN locations and Engagement Metrics

The column indicating the VPN country used during data collection was removed, as it was only relevant for dataset merging during the collection phase (Section 3.1) and not required for the analysis. In addition, three new engagement-related metrics were introduced: like/view ratio, comment/view ratio, and share/view ratio (the latter specific to TikTok only). These ratios were added to account for the large variation in view counts across individual videos. By normalizing engagement through these relative measures, comparisons between videos became more meaningful when evaluating the potential impact of AI labeling on user interaction.

Some TikTok videos contained invalid values for the `like_view_ratio` and `comment_view_ratio`. These cases were manually validated. Videos that genuinely had no views, likes, or comments were removed from the dataset, while others were corrected based on the available metadata.

3.2. Data Sampling for Sub-Question 4

To assess labeling consistency for Sub-Question 4, a dataset was required that included both the actual label status and an assessment of whether a label would have been appropriate. The information about the actual label status was gathered from the full dataset collected in the previous section. This sample was then reviewed for the label appropriateness, further explained in Section 4.2. First, 1% of the data was reviewed by two researchers, and after reaching an appropriate level of agreement, a total of 5% was reviewed by one coder.

3.2.1. Data Sampling

An initial sample of 140 videos (slightly over 1% of the full dataset collected in the last section) was selected to assess inter-rater reliability. This number allowed for an equal distribution across four groups of 35 videos each. To account for potential video removals between data collection and sampling, a larger initial pool of 180 videos—45 per group—was drawn. In the case of TikTok, the labeled group included both creator-applied and platform-applied labels.

After reaching an appropriate level of agreement based on the inter-rater reliability analysis, the full sample of 5% was reviewed by one coder. As with the initial round, a slightly larger pool was drawn in advance to account for potential video removals or privacy changes.

When expanding the sample to approximately 5% of the full dataset, the earlier 35 reviewed videos were retained. To avoid duplicate coding, the entire set of 45 initially sampled videos per group, including those not ultimately coded, was excluded from the second sampling round. From the remaining pool, 109 additional videos per group were randomly selected using stratified sampling, resulting in a total of 154 videos per group. One coder labeled the additional videos. Table 3.3 presents the distribution of all reviewed videos, combining both coding rounds into a single overview.

Table 3.3: Overview of Reviewed Videos by Platform, AI Label Status, and Coding Round (First vs. Second Sample)

Platform	**AI Label**	**First Sample**		**Second Sample**		**Total Videos Reviewed**
		# in sample	# reviewed (2 coders)	# in sample	# reviewed (1 coder)	
YouTube	Without AI label	45	35	140	119	154
YouTube	With AI label	45	35	140	119	154
TikTok	Without AI label	45	35	140	119	154
TikTok	With AI label	45	35	140	119	154
Total						**616**

3.2.2. Data Cleaning

After each reviewing round, the corresponding dataset was cleaned. Several steps were taken to prepare the data for analysis:

- The Excel sheet was manually adjusted to ensure that only completed coding rows were included.
- Datasets were split by platform (YouTube and TikTok), and columns with irrelevant suffixes (e.g., ‘.1‘, ‘.2‘) or unnamed placeholders were removed.
- Videos flagged as unavailable (e.g., deleted or set to private) were excluded from further analysis.
- Consistency checks ensured that subcategories were only assigned when the video was marked as AI-generated, and vice versa. No mismatches were found during this process.
- The two coding files were aligned based on shared video IDs. Some rows appeared in only one of the two files, due to video removals between the time of coding. These unmatched cases were excluded to ensure a fair and complete comparison.

After each cleaning round, only those video entries that were (a) available on the platform, (b) present in both coding files, and (c) internally consistent, were retained. The cleaned dataset from the first sample was used for the inter-rater reliability analysis, and the final cleaned dataset after the second sample expansion served as the basis for the labeling consistency analysis. Both analyses are presented in the next chapter.

4

Methodology

Following the data collection and cleaning described in the previous chapter, this chapter outlines the methodology applied to address Sub-Questions 3 and 4.

4.1. Methodology for Sub-Question 3

The goal of Sub-Question 3 is to assess how prevalent AI labels are and what their potential impact is on user engagement metrics. Section 2.3 revealed clear differences in how TikTok and YouTube apply these labels. On TikTok, the user can see whether labels are applied by the creators themselves or by the platform. In contrast, YouTube does not distinguish between creator- and platform-applied labels, making it impossible to determine the source of each label. Additionally, YouTube often does not display the label within the video player, whereas TikTok does.

This section outlines the analytical approach used to examine the relationship between AI label presence and user engagement metrics within each platform. By analyzing labeled and non-labeled videos separately for TikTok and YouTube Shorts, the methodology enables a platform-specific assessment of potential differences in engagement outcomes.

4.1.1. Platform specific Approach

Given the differences in how TikTok and YouTube implement and display AI labels, a platform-specific analytical approach was adopted to examine their relationship with user engagement.

TikTok

For TikTok, the potential impact of AI labels on engagement metrics (likes, comments, shares, views) was analyzed using the creator-applied labels only. These labels are voluntarily added by creators at the time of upload and cannot be altered afterwards. This provides a clearer basis for interpreting their potential influence on engagement, as the label is known to be present when users first interact with the content.

YouTube Shorts

For YouTube Shorts, it is not possible to determine whether AI labels were applied by the platform or the content creator, nor whether they were present at the time of user engagement. As a result, the analysis is limited to comparing engagement metrics between labeled and non-labeled videos, without drawing any causal conclusions.

It is also important to keep in mind that YouTube's AI labels are typically only visible in the expanded video description. An exception is made for sensitive content, where the label appears directly in the video player (see Section 2.3). In this dataset, only 0.4% of videos were flagged as sensitive. When considering all AI-labeled videos, only 0.8% featured a label that was visible without user interaction. This further limits the interpretability of any observed differences in engagement, as users are unlikely to be aware of the label in most cases.

4.1.2. Engagement Analysis Approach

It is essential to select an appropriate statistical approach to assess the potential relationship between AI labels and user engagement. Many statistical tests assume that the data is normally distributed. However, as shown in the exploratory data analysis (Figure B.1), all engagement metrics are highly skewed. Visual inspection suggests that the distributions of engagement metrics are similar in shape for both AI-labeled and non-AI-labeled videos.

Given this non-normality, the Mann–Whitney U test was selected to test the potential relationship between AI labels and engagement metrics. This is a non-paramatric test which does not assume normality and is suitable for skewed distributions [85]. With this test, it is possible to compare two independent groups [17]. This aligns with the structure of this study: the independent variable (label presence) divides the sample into two distinct, non-overlapping groups (labeled and non-labeled videos). A significance level of $p < 0.05$ was applied in line with conventional statistical standards [3]. Although the use of this threshold has been debated, it remains generally accepted for exploratory research and hypothesis testing. Exact p-values are also reported for a more nuanced interpretation of the results.

Literature emphasizes that p-values alone are not sufficient for interpreting results [28, 84]. Therefore, the standardized effect size `r` was also calculated (see Equation 4.1). As discussed by Pautz, Olivier, and Steyn [66], the equation was originally introduced by Rosenthal [71].

Reporting effect sizes is particularly important, as Farmus et al. [28] found that 93% of reported values were not interpreted using standard thresholds. This study follows the guidelines of Cohen [16], which define effect sizes as follows:"

- r = 0.1 to 0.3: Small effect
- r = 0.3 to 0.5: Medium effect
- r > 0.5: Large effect

In Python, the `mannwhitneyu` function from the `scipy.stats` package was used to perform the Mann-Whitney U test. This results in both the U-statistic and the p-value.

The formula shown in Equation 4.1 was utilized to calculate the effect size r. In this formula, n_a and n_b indicate the sizes of the samples of the two independent groups. The U-statistic is standardized to a Z-score using its expected value m_U and standard deviation s_U, as defined in the same equation. The standardized Z is then divided by $\sqrt{n_a + n_b}$ to obtain the effect size r:

$$r = \frac{Z}{\sqrt{n_a + n_b}}$$
$$\text{where } Z = \frac{U - m_U}{s_U}, \quad m_U = \frac{n_a n_b}{2}, \quad s_U = \sqrt{\frac{n_a n_b (n_a + n_b + 1)}{12}} \tag{4.1}$$

As discussed in the literature review (Section 2.3), previous findings of the impact of AI content labels are mixed. Some studies report reduced engagement due to lower perceived credibility, while others observe increased interaction driven by user curiosity. Given these conflicting results, this study does not assume a specific direction of effect. Instead, it tests whether the presence of a label is associated with any statistically significant difference in user engagement.

As a result, multiple hypotheses were tested in this study. This was done for each engagement metric (like/view, comment/view, share/view and views). The general null hypothesis for each test was that there is no significant difference in median engagement between labeled and non-labeled videos. A full overview of these hypotheses can be found in Appendix D.

The methodology described in this chapter provides a robust framework for evaluating whether the presence of AI content labels is associated with differences in user engagement. In the following section. the results of this analysis are presented.

4.2. Methodology for Sub-Question 4

The goal of Sub-Question 4 is to evaluate how consistently AI labels are applied in practice, and whether this application aligns with platform-specific guidelines. As outlined in Section 2.3, TikTok and YouTube Shorts differ not only in how they display AI labels, but also in how their documentation defines when such labels are required. This raises concerns about the clarity and enforceability of these policies.

This section presents the methodological approach used to assess labeling consistency on both platforms. It includes a structured thematic coding procedure, inter-rater reliability analysis, and a classification-based comparison of expected versus observed label presence. These steps allow for a systematic evaluation of whether AI labels are applied in accordance with each platform's own rules.

4.2.1. Thematic Coding

Determining whether a video should have an AI label is subjective: it requires interpreting both the video content and the platform's often vaguely formulated guidelines. Adding to the complexity, TikTok and YouTube differ in both the definition and enforcement of AI labeling.

To reduce subjectivity and ensure a structured basis for evaluation, a deductive thematic coding approach was adopted. Thematic coding is a qualitative technique for detecting and organizing recurring patterns within a dataset [13].

To limit interpretive variance, the coding process was guided by a predefined codebook derived from official platform documentation [30, 89, 93, 101, 103, 104]. This deductive (top-down) approach relies on predefined categories to classify content, in contrast to inductive methods where categories emerge from the data itself [70]. The main code categories are summarized in Table 4.1, and the full codebook is provided in Appendix C. To ensure clarity and consistency, the codebook was reviewed by two researchers prior to implementation. The sensitive topic categorization is only included for YouTube, since this influences the location of the AI label.

Table 4.1: Codebook Main Categories

Code (YouTube)	**Description**
Y1	AI-generated Content
Y2	Sensitive Topic
Code (TikTok)	**Description**
T1	AI-generated Content

The sample of 140 videos, described in Table 3.3, was independently coded by two researchers. Each coder first determined whether the content was AI-generated[1]. If the content was AI-generated, the coder then assigned the appropriate category. Finally, the coders noted the presence of sensitive content for YouTube Shorts.

Based on these assigned subcategories and the official platform documentation, a derived variable was added to indicate whether an AI label was required for each video. This labeling decision was not made manually by the coders; instead, it followed directly from the applied categories. The specific subcategories that, according to each platform's policy, require AI labeling are listed in Appendix C.

4.2.2. Inter-rater Reliability

To assess the reliability of the coding process, inter-rater reliability (IRR) was calculated. IRR reflects the degree of agreement between coders analyzing the same content. According to Gisev, Bell, and Chen [34], selecting a suitable IRR statistic depends on the characteristics of the data. Given that the coding categories were nominal and two coders were involved, Cohen's Kappa was selected as the most suitable statistic. Although percentage agreement is a commonly reported IRR measure [57], it does not account for agreement occurring by chance. To account for chance agreement, Cohen's Kappa offers a more robust estimate. The formula for Cohen's Kappa is provided below [57]:

$$\kappa = \frac{p_a - p_e}{1 - p_e} \tag{4.2}$$

In Equation 4.2, p_a denotes the observed agreement, while p_e represents the expected agreement by chance. The calculation was performed in Python using the `cohen_kappa_score` function from the `sklearn.metrics` module.

To assess the reliability of the coding process, inter-rater reliability (IRR) was calculated for two variables:

[1] On TikTok, it is referred to as AI-generated content, while on YouTube it is described as synthetic or altered content. For clarity, only AI-generated content is used throughout this thesis to refer to both.

1. The classification of AI-generated content.
2. The derived column indicating whether a label was appropriate.

Since the research question focuses on the (in)consistency of label application, the second column was treated as most relevant. Following Landis and Koch [48], a threshold of $\kappa \geq 0.61$ was considered to reflect substantial agreement. Agreement on this derived "label appropriate" column was treated as a condition for using the data in the next stage of analysis.

Disagreements on whether a label was required were discussed between the two coders. After consensus was reached, the subcategories were adjusted where necessary. One coder then proceeded to label the remainder of the dataset, amounting to approximately 5% of the total data.

4.2.3. Assessing Labeling Consistency

Following the resolution of disagreements and completion of coding, the full dataset was analyzed to assess whether AI labels were applied in accordance with platform-specific guidelines. For each video, the actual label status was compared to the expected status derived from the assigned subcategories. This allowed classification into four categories:

1. True Positives (TP): Videos with an AI label, where a label was appropriate.
2. False Positives (FP): Videos with an AI label, where no label was appropriate.
3. True Negatives (TN): Videos without an AI label, where no label was appropriate.
4. False Negatives (FN): Videos without an AI label, where a label was appropriate.

These categories, summarized in Table 4.2, form the basis for evaluating the consistency of AI label application across platforms.

Table 4.2: Confusion matrix used to assess consistency of AI label application

	AI Label Appropriate	**AI Label Not Appropriate**
AI Label Applied	True Positive (TP)	False Positive (FP)
No AI Label Applied	False Negative (FN)	True Negative (TN)

Because each video is assigned both an expected label status (required or not required) and an actual observed status (present or absent), the task can be treated as a binary classification problem. This enables the use of standard performance metrics [55]. Following Marioriyad and Ramazi [55], these metrics are defined in Equation 4.3 and can be used to assess the consistency of AI label application relative to platform guidelines.

$$\begin{aligned} \textit{accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} & \textit{specificity} &= \frac{TN}{TN+FP} \\ \textit{precision} &= \frac{TP}{TP+FP} & \textit{recall} &= \frac{TP}{TP+FN} \end{aligned} \tag{4.3}$$

Each metric captures a different aspect of labeling consistency:

1. *Accuracy:* The proportion of videos for which the AI label status is correctly applied, whether it is present when required or absent when not.
2. *Specificity:* The proportion of videos without AI-generated content that are correctly not labeled.
3. *Precision:* The proportion of labeled videos that actually contain AI-generated content, indicating how often a present label is appropriate.
4. *Recall:* The proportion of AI-generated videos that are correctly labeled, reflecting how often required labels are actually applied.

The methodologies outlined in this chapter provide a structured approach to evaluate how accurately and consistently AI labels are applied to short-form video platforms. Through thematic categorization based on platform documentation, an expected label status was assigned to each video. Afterwards, this was compared to the observed status. By applying classification metrics to these comparisons,

the methodology allows for a quantifiable assessment of labeling practices relative to platform-specific guidelines. The following chapter outlines the analysis results and compares TikTok and YouTube Shorts in terms of their similarities and differences.

5 Results

This chapter presents the results of the data analyses described in Chapter 4. Section 5.2 addresses Sub-Question 3, focusing on the prevalence of AI labels and their potential relationship with user engagement on TikTok and YouTube Shorts. Section 5.3 presents the results for Sub-Question 4, examining the consistency of AI labeling practices across both platforms.

5.1. Exploratory Data Analysis

This section examines the structure and key features of the dataset to provide initial insights into the use of AI labels and platform-specific engagement patterns. These findings inform and contextualize subsequent statistical analyses.

The data collection process described in the previous chapter resulted in a dataset of 13,163 videos, including 7,082 from TikTok and 6,071 from YouTube Shorts.

5.1.1. Filtering by Time Range

To ensure meaningful comparisons between TikTok and YouTube Shorts, the dataset was filtered to align the time ranges for both platforms. The YouTube dataset initially contained data from January 2022 onward, while the TikTok dataset included data starting from January 2020. To explore how AI label application evolved over time, the number of videos with and without labels was plotted for each platform in Figure 5.1 and Figure 5.2.

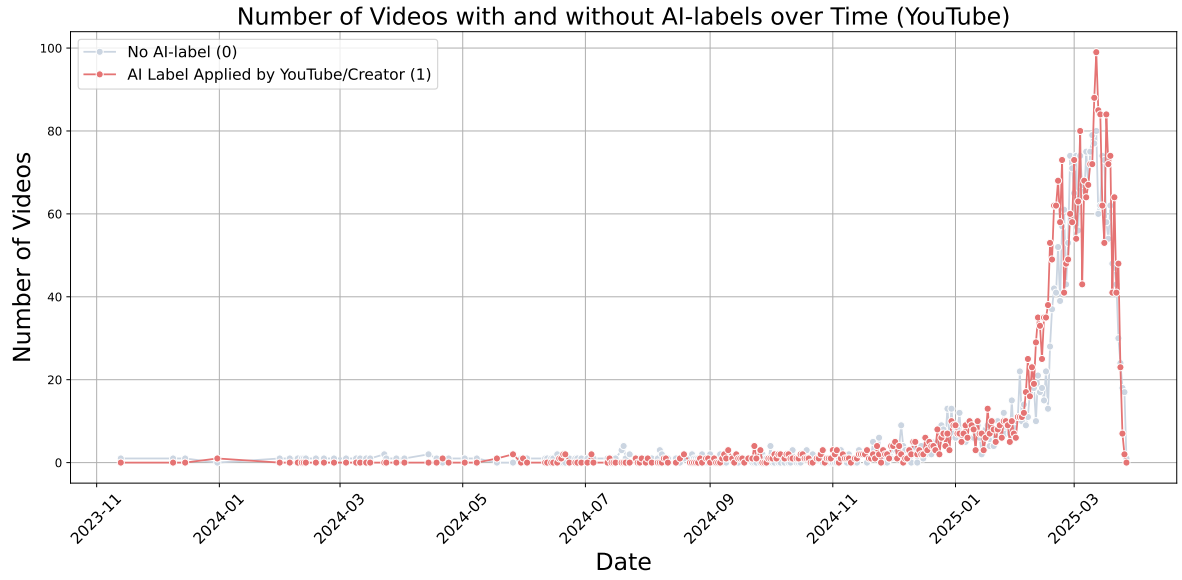


Figure 5.1: AI Label Application on YouTube Shorts over Time

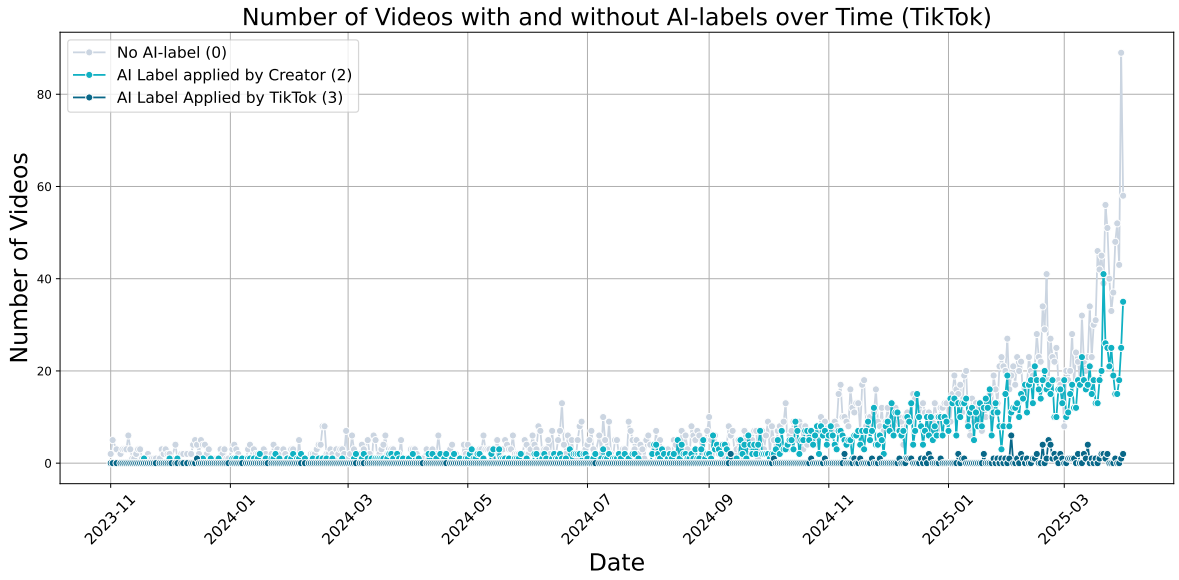


Figure 5.2: AI Label Application on TikTok over Time

As shown in the figure, almost no videos were labeled prior to 2023. Upon further investigation, this was attributed to the fact that AI content labels were only introduced in September 2023 on TikTok [90] and in November 2023 on YouTube [30]. Consequently, the dataset was filtered to include only videos published from November 2023 onward for both platforms.

As shown in Figure 5.2, a notable observation is that the majority of AI-labeled TikTok videos appear to have been labeled by the creators themselves. In contrast, YouTube does not distinguish between labels applied by the platform and those applied by the creator.

Table 3.2 provides an overview of the available metadata columns per platform. Platform-specific features—such as TikTok's `shares` and `saved`, and YouTube's `sensitive_topic`—were retained for within-platform analyses.

5.1.2. Hashtag Usage per Platform

To better understand the differences in AI-related hashtag usage between TikTok and YouTube Shorts, the distribution of hashtags in the collected datasets was analyzed. As described in Section 3.1, the hashtag set was constructed iteratively. It started from the core hashtag #ai and was expanded with frequently co-occurring, AI-related hashtags identified during the scraping process.
Figure 5.3 shows the percentage of videos per platform that include each AI-related hashtag from the hashtagset (Table 3.1. The figure highlights two key observations:

- On both platforms, a substantial share of videos included the hashtag #ai. This suggests that using #ai as the core hashtag was a reasonable starting point for capturing AI-related content.
- YouTube Shorts appears to use a broader variety of AI-related hashtags than TikTok. This suggests that YouTube videos are more likely to include multiple AI-related hashtags within a single title or description.

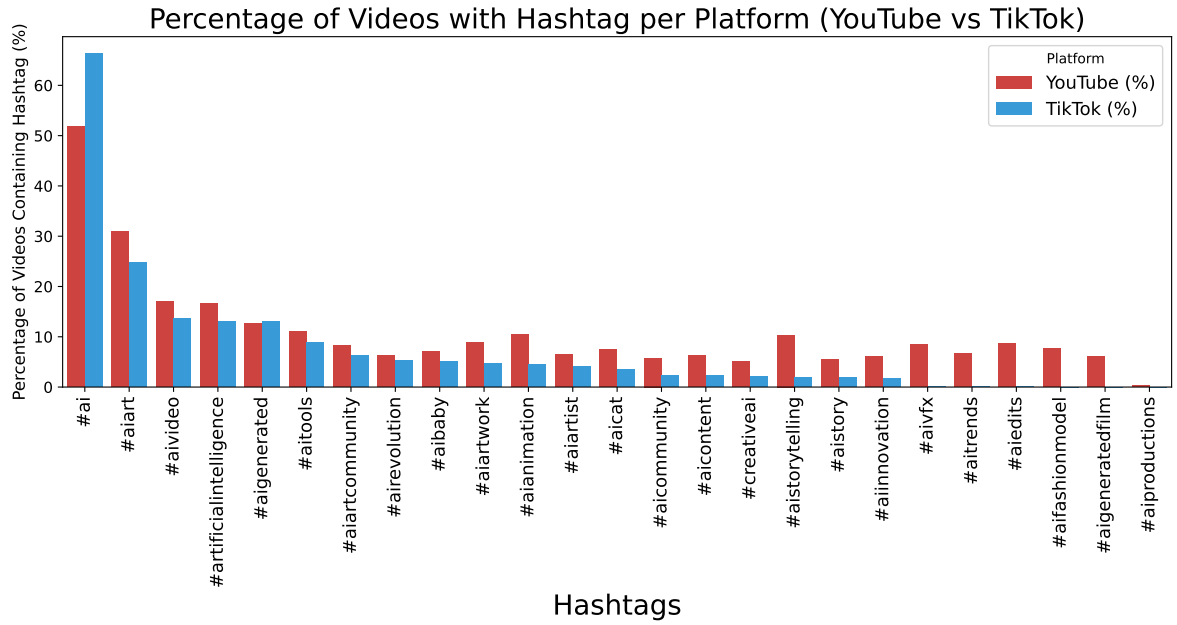


Figure 5.3: Percentage of Videos with Hashtag per Platform (YouTube Shorts vs. TikTok)

5.1.3. Engagement Metrics

Before examining the potential impact of AI labels on engagement, the data distribution must first be understood. As previously mentioned, three variables were added to the dataset: `like/view ratio`, `comment/view ratio`, `views` and `share/view ratio` (the latter specific to TikTok).

During data validation, it was found that the `like/view ratio` on TikTok occasionally exceeded 1. Upon closer inspection, two videos had more likes than reported views. In one case, the discrepancy could not be corrected because the engagement metrics had increased so much between the original data collection and the manual validation that the original values could no longer be verified. In the other case, the video genuinely had more likes than views, likely due to a platform bug or inconsistency in how TikTok tracks views. As these values were not considered reliable, both rows were removed from the dataset.

After cleaning the engagement metrics, all the distributions were found to be highly left-skewed, as shown in Figure B.1.

5.2. AI Label Prevalence and Impact on Engagement (SQ3)

This section presents the results addressing Sub-Question 3, which focuses on the prevalence of AI labels and their potential relationship with user engagement on TikTok and YouTube Shorts.

5.2.1. Prevalence of AI Labels

To explore how AI labels are used across platforms, the proportion of labeled and non-labeled videos was visualized in Figure 5.4 and Table 5.1.

The results show a clear difference in the absence of AI labels between platforms. On TikTok, the majority of videos (63.13%) did not include an AI label, compared to 48.10% of the YouTube Shorts. Additionally, as discussed in Section 3.1.3, YouTube does not differentiate between label sources. On TikTok, most AI labels were applied by creators (35.34%), and only a small proportion by the platform itself (1.53%). When combining both creator- and platform-applied labels to enable cross-platform comparison, 36.87% of TikTok videos contained an AI label, which remains lower than the prevalence observed on YouTube Shorts (51.90%). It was also found that only 25 videos (0.41% of the total YouTube Shorts dataset) had a “sensitive topic” label, indicating that such labels are very rare. These descriptive findings set the stage for further analysis of how label presence might relate to user engagement.

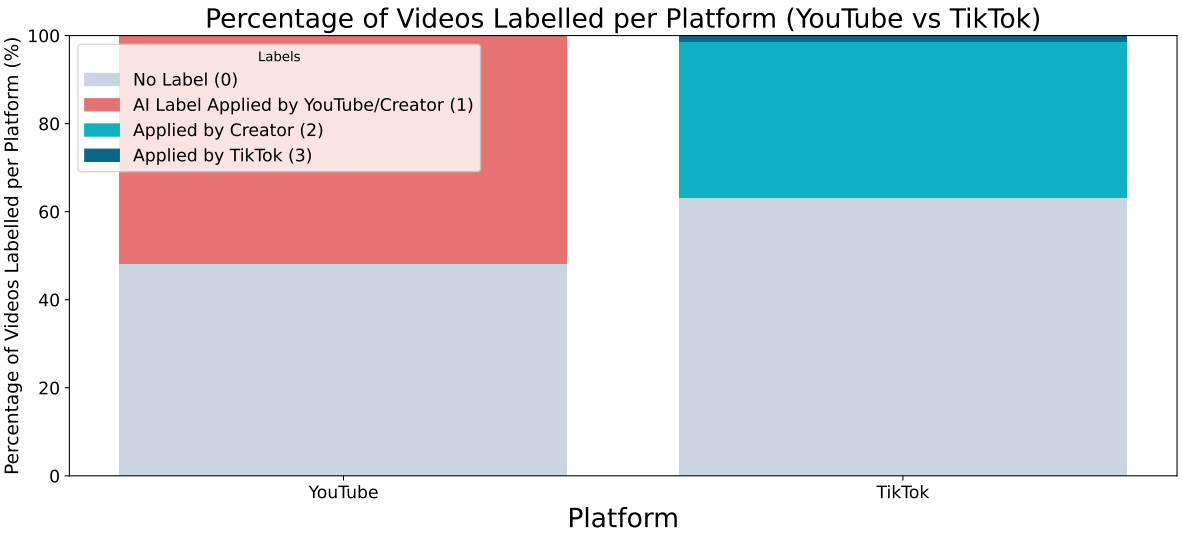


Figure 5.4: Percentage of Labeled vs. Non-Labeled Videos per Platform

Table 5.1: Distribution of AI Labels by Platform (in %)

Label Status	YouTube Shorts(%)	TikTok (%)
No Label (0)	48.10%	63.13%
Applied by YouTube/Creator (1)	51.90%	N/A
Label Applied by Creator (2)	N/A	35.34%
Label Applied by TikTok (3)	N/A	1.53%

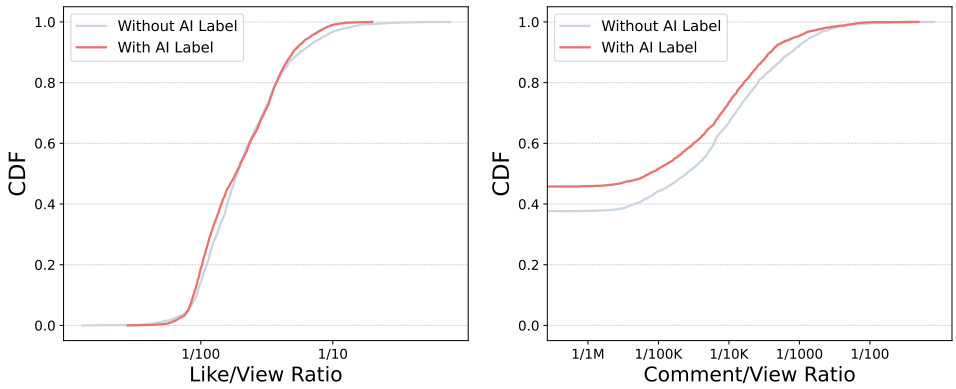
5.2.2. Impact of AI Labels on Engagement Metrics

To examine possible differences in engagement associated with AI labels, cumulative distribution plots were first inspected. Subsequently, the Mann-Whitney U test was conducted as described in Section 4.1. This determined whether these observed differences are statistically significant. The tested hypotheses are listed in Appendix D.

YouTube Shorts

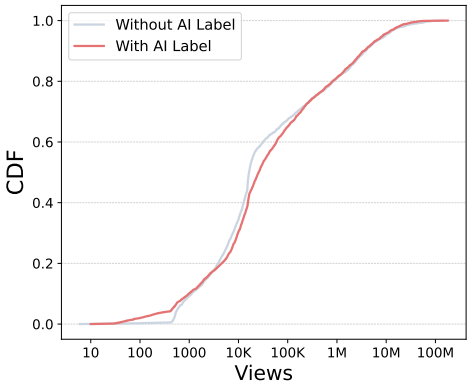
The CDF plots in Figure 5.5 show differences in the distributions of the like/view ratio (Figure 5.5a) and the comment/view ratio (Figure 5.5b). For both metrics, higher ratios are observed for videos without an AI label across most of the distribution. In the case of views (Figure 5.5c), the CDF curves show that videos without an AI label are more common at lower view counts (below approximately 1,000 views). Videos with an AI label are more frequent between 1,000 and 1,000,000 views. For very high view counts, the distributions converge again.

The Mann-Whitney U test results (Table 5.2) show that the null hypothesis is rejected for all engagement metrics ($\alpha < 0.05$). This indicates statistically significant differences between the groups. The median like/view and comment/view ratios are lower for videos with an AI label, indicating a higher probability that a randomly selected video without an AI label has a higher like/view or comment/view ratio than a video with an AI label. However, according to the interpretation guidelines of Cohen [16], the effect sizes for all metrics remain below the threshold for a small effect and are therefore considered negligible.



(a) Like/View Ratio

(b) Comment/View Ratio



(c) Views

Figure 5.5: Cumulative Distribution Functions of Engagement Metrics for YouTube Videos

Table 5.2: Summary Statistics YouTube (significance level $\alpha < 0.05$)

Metric	Mann-Whitney U	p-value	Effect size (r)	Direction (median)	H_0 Rejected?
Like/View Ratio	4744082	0.00	0.04	Label < No Label	Yes
Comment/View Ratio	5027525	0.00	0.09	Label < No Label	Yes
Views	4342571	0.00	0.04	Label > No Label	Yes

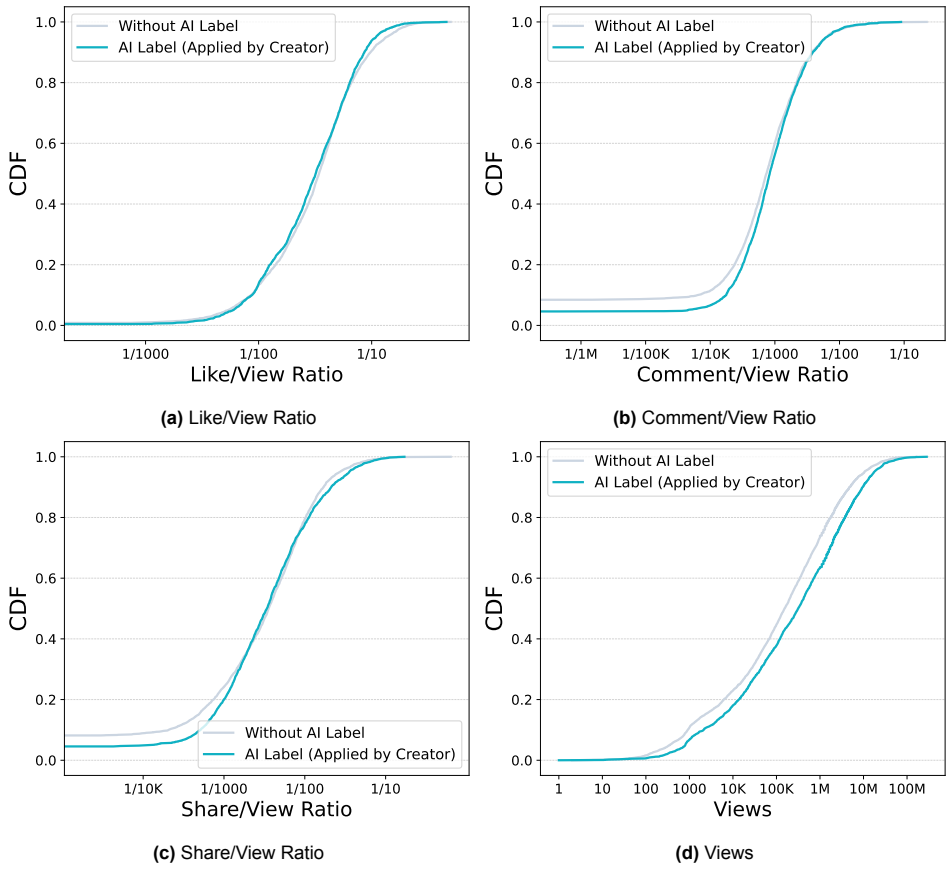
TikTok

The CDF plot in Figure 5.6a shows minimal differences in the like/view ratio distribution between videos with and without an AI label. Where differences occur, videos without an AI label tend to have slightly higher ratios. In contrast, the comment/view ratio (Figure 5.6b) is generally higher for videos with an AI label across most of the distribution. A similar, though less pronounced, pattern is observed for the share/view ratio (Figure 5.6c). For views (Figure 5.6d), videos with an AI label consistently achieve higher view counts throughout the distribution.

The Mann-Whitney U test results (Table 5.3) show that the null hypothesis is rejected for the like/view ratio, comment/view ratio, and views metrics ($\alpha < 0.05$), indicating statistically significant differences between the groups for these metrics. No significant difference was found for the share/view ratio.

Regarding the direction of the differences, the results indicate a higher probability that a randomly selected video without an AI label has a higher like/view ratio than a labeled video. For the comment/view ratio and views, the opposite is observed: a randomly selected video with an AI label is more likely to have a higher comment/view ratio and view count than a video without an AI label.

According to the interpretation guidelines in Section 4.1, the effect sizes for the like/view ratio, comment/view ratio, and share/view ratio are below the threshold for a small effect and are therefore considered negligible. Only the effect size for views reaches the threshold for a small effect.



(a) Like/View Ratio

(b) Comment/View Ratio

(c) Share/View Ratio

(d) Views

Figure 5.6: Cumulative Distribution Functions of Engagement Metrics for TikTok Videos

Table 5.3: Summary Statistics TikTok (significance level $\alpha < 0.05$)

Metric	Mann-Whitney U	p-value	Effect size (r)	Direction (median)	H_0 Rejected?
Like/View Ratio	4534285	0.03	0.03	Label < No Label	Yes
Comment/View Ratio	4099186	0.00	0.06	Label > No Label	Yes
Share/View Ratio	4292708	0.14	0.02	Label < No Label	No
Views	3835207	0.00	0.11	Label > No Label	Yes

5.3. Assessing AI Labeling Consistency (SQ4)

To address Sub-Question 4, this section presents the outcomes of the labeling consistency analysis. It evaluates whether AI labels were applied in accordance with platform guidelines, based on a coded sample of TikTok and YouTube Shorts videos. The results provide insight into the frequency of both overlabeling and underlabeling across platforms.

5.3.1. Inter-rater Reliability

As described in Section 4.2, Cohen's Kappa was applied to evaluate the reliability of the coding process. Following the first round of coding, agreement was calculated for two variables: (1) the classification of AI-generated content and (2) the derived column indicating whether an AI label was appropriate.

Table 5.4 presents the resulting scores. Both values fall within the range that Landis and Koch [48] interpret as indicating substantial agreement ($\kappa \geq 0.61$). These results confirmed that the coding scheme was sufficiently reliable to proceed with the remaining dataset. Based on this outcome, one coder completed the annotation of the full 5% sample used for the consistency analysis.

Table 5.4: Cohen's Kappa Scores for Inter-Rater Reliability on AI Category Classification and Label Appropriateness

Category of Agreement	Cohen's Kappa
Category	0.68
Label	0.69

5.3.2. Labeling Consistency

Each video in the sample outlined in Table 3.3 was evaluated by comparing its actual label status to the label status that was deemed appropriate based on the assigned subcategories and platform policies.

The results are interpreted using the binary classification framework introduced in Table 4.2, which distinguishes between four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

The performance metrics precision, recall, accuracy, and specificity were calculated for each platform to evaluate labeling consistency. These metrics, defined in Equation 4.3, provide a structured basis for comparing actual versus expected label application.

Table 5.5: Labelling Consistency YouTube

Platform	Metric	Count	%
YouTube	True Positives (TP)	49	15.91%
	False Positives (FP)	105	34.09%
	True Negatives (TN)	125	40.58%
	False Negatives (FN)	29	9.42%
	Total	**308**	**100%**

Table 5.6: Labelling Consistency TikTok

Platform	Metric	Count	%
TikTok	True Positives (TP)	151	49.03%
	False Positives (FP)	3	0.97%
	True Negatives (TN)	80	25.97%
	False Negatives (FN)	74	24.03%
	Total	**308**	**100%**

YouTube Shorts

Table 5.5 presents the absolute and relative frequencies of correct and incorrect label applications on YouTube Shorts:

- True Positives (TP): In 15.91% of cases, an AI label was correctly applied.
- False Positives (FP): In 34.09% of cases, an AI label was applied when it was not required, indicating a strong tendency toward overlabeling.
- True Negatives (TN): In 40.58% of cases, no AI label was applied, and this was correct. This represents the most frequent outcome.
- False Negatives (FN): In 9.42% of cases, no AI label was applied when one was actually required, pointing to a smaller but noticeable degree of underlabeling.

Overall, the results suggest that overlabeling occurs more frequently than underlabeling, and that correctly not applying a label is the most consistent outcome.

TikTok

Table 5.6 shows the distribution of label consistency outcomes for TikTok videos:

- True Positives (TP): In 49.03% of cases, an AI label was applied correctly.
- False Positives (FP): Only 0.97% of videos were incorrectly labeled as AI-generated when a label was not required, suggesting minimal overlabeling.
- True Negatives (TN): In 25.97% of cases, no label was applied and this was appropriate.
- False Negatives (FN): In 24.03% of cases, a required AI label was missing, indicating some degree of underlabeling.

These results suggest that correct labeling is most often achieved when a label is required, and incorrect labeling is relatively rare. However, about a quarter (24.03%) of TikTok videos remain unlabeled despite requiring an AI label, suggesting that AI labels are still not applied frequently enough.

Performance Metrics

Table 5.7 presents the performance metrics for both platforms. For YouTube, the relatively low precision (0.3180) suggests that many applied labels were unnecessary, consistent with the observed overlabeling. The recall (0.6280) shows that approximately 63% of videos requiring a label were correctly labeled. While the accuracy (0.5650) and specificity (0.5430) indicate moderate overall labeling performance, this means that just over half of all videos were labeled correctly (accuracy), and only about 54% of videos without AI-generated content correctly received no label (specificity).

In contrast, TikTok achieves a very high precision (0.9810), indicating that nearly all applied labels were appropriate. The recall (0.6710) shows that most, but not all, required labels were correctly applied. TikTok also demonstrates higher accuracy (0.7500) and specificity (0.9640), reflecting more consistent and precise labeling outcomes compared to YouTube. An accuracy of 75% means that three out of four videos were labeled correctly, regardless of whether a label was required. The high specificity of 96.40% further indicates that in most cases, videos that do not require an AI label remain unlabeled.

Table 5.7: Performance metrics thematic coding

Platform	Precision	Recall	Accuracy	Specificity
YouTube	0.3180	0.6280	0.5650	0.5430
TikTok	0.9810	0.6710	0.7500	0.9640

Types of wrongly classified videos

False Positives

For YouTube, 99% of the false positives (105 out of 106) were classified as *Y1.5: Clearly unrealistic content*. According to YouTube's guidelines, as shown in Appendix C, this type of content does not require an AI label. The remaining false positive was classified as *Y1.7: Adding elements to the video*. This video depicted a human digestive system and was also the only video in the entire YouTube sample (1 out of 308) classified as sensitive content.

For TikTok, only three false positives were identified. Two involved videos classified as *Y1.7: Effects/Filters*, where AI filters were applied to images or videos, or synthetic voices in another language were added. TikTok's guidelines do not require a label for such content. The remaining case was a video explaining how to create a storyboard using AI tools. Based on the coding guidelines agreed during thematic analysis, this type of instructional content was not considered AI-generated, as it did not involve content directly created by AI.

False Negatives

False negatives were more common on both platforms. On TikTok, 85% of the 74 false negatives were classified as *Clearly unrealistic content (T1.5)*. Unlike YouTube, TikTok requires creators to label this type of content, yet labels were still frequently missing.

The remaining cases on both platforms involved similar types of content:

- *Using the likeness of a realistic person (T1.1/Y1.1):* Celebrities were altered to perform unrealistic actions, such as dancing or heroic acts. Although this content often appeared animated and clearly unrealistic, it was classified under this category because it featured recognizable public figures.
- *Altering footage of real events or places (T1.2/Y1.2):* Content using real event footage, such as talent shows, with unrealistic elements digitally added. Examples include contestants appearing to transform from lions into humans, people or vehicles compressed like a sponge, or dogs appearing to fly.
- *Generating realistic scenes (T1.3/Y1.3):* Content featuring highly realistic but non-famous individuals, making it difficult to assess whether it was AI-generated. This also included "spot the AI" challenges and surreal scenes, such as creatures attacking boats.

Classification of AI-Generated Content

Figure 5.7 presents the percentage of videos classified as AI-generated content on each platform. These values reflect the share of videos identified as AI-generated during the thematic coding pro-

cess. It is important to note that this figure does not consider whether the application of an AI label was appropriate according to the platforms' guidelines.

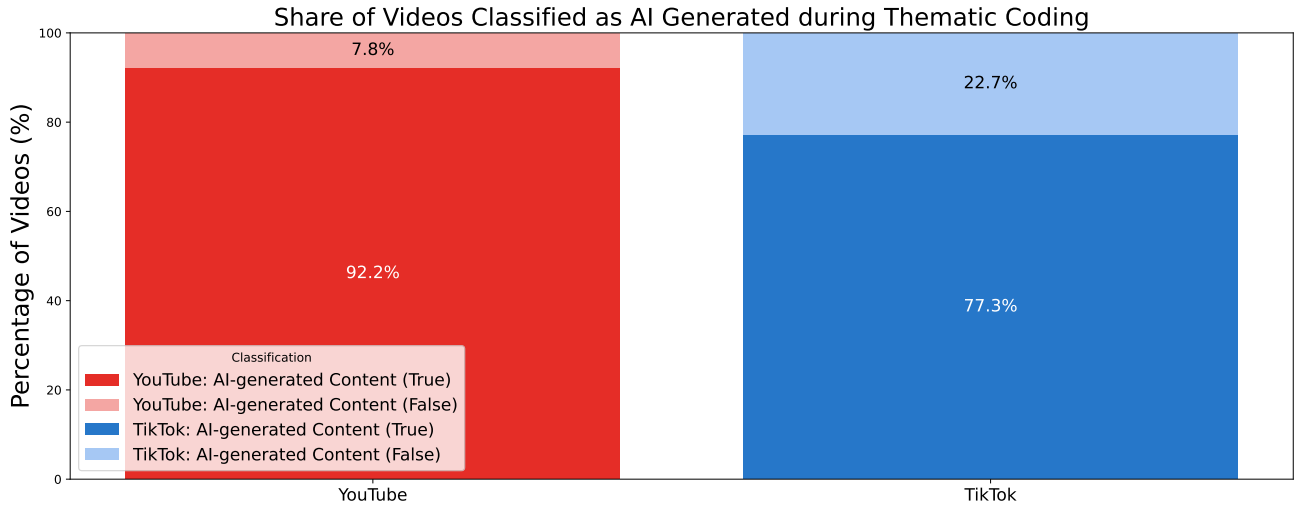


Figure 5.7: Share of videos classified as AI-generated per platform, based on thematic coding

5.4. Key Takeaways

The exploratory and statistical analyses reveal notable differences between YouTube Shorts and TikTok:

- AI Label Prevalence: AI labels are more frequently applied on YouTube Shorts (51.90%) than on TikTok (36.87%).
  - On YouTube, only 0.41% of these labels appear directly in the player; 51.49% are shown in the expanded description.
  - On TikTok, 35.34% of labels were creator-applied, and only 1.53% were applied by the platform.
- Impact on Engagement: While statistically significant differences were found, effect sizes remain negligible for most metrics.
  - Both platforms show lower like/view ratios for labeled videos.
  - Comment/view ratios show mixed results: lower for labeled videos on YouTube, higher on TikTok.
  - Only the views metric on TikTok showed a small effect size ($r > 0.1$); indicating a higher probability that a randomly selected labeled video has more views than a non-labeled video ($\alpha = 0.00$).
- Labeling Consistency: TikTok applies AI labels more accurately (Precision: 98.10%; Accuracy: 75.00%) than YouTube Shorts (Precision: 31.80%; Accuracy: 56.60%).
  - YouTube Shorts exhibits frequent overlabeling (false positives).
  - TikTok shows more underlabeling (false negatives).

Overall, these findings suggest that although AI labels are increasingly present, their application varies significantly between platforms in terms of visibility and consistency. In the next chapter, these results will be further interpreted and discussed.

6 Discussion

This chapter reflects on the findings presented in the previous chapter and discusses them in the context of the existing literature and platform guidelines. The goal is to interpret the results beyond statistical significance and performance metrics, exploring the practical implications for platform governance and policy development. Additionally, this chapter addresses the limitation of the research and outlines recommendations for platforms, policy makers and future studies.

6.1. Summary of Key Findings

This section summarizes the key findings and relates them to insights from the existing literature.

6.1.1. Prevalence of AI Labels

Within the dataset of this study, AI labels were found to be more prevalent on YouTube Shorts than on TikTok. More than half of the YouTube Shorts included an AI label (51.9% of the 6,039 videos), whereas only 37.8% of the 6,279 TikTok videos were labeled. Of the TikTok labels, just 1.5% were applied by the platform itself, indicating that labeling on TikTok is more creator-driven.

As discussed in Chapter 2, both platforms use the C2PA standard to detect and apply AI labels. Despite this shared technical approach, the prevalence of AI labels differs substantially. One possible explanation is that the nature of the content differs between platforms. Figure 5.7 shows that a larger share of YouTube Shorts was classified as AI-generated compared to TikTok during the thematic coding. This may suggest that TikTok users produce less AI-generated content while using AI-generated hashtags, which could partially account for the lower labeling rate. Additionally, thematic coding revealed that TikTok creators tend to underlabel their content. The combination of a potentially lower share of AI-generated content and a higher degree of underlabeling may explain the lower prevalence of AI labels on TikTok. However, it is important to note that these observations are based on a sample of the data and may not fully represent the entire dataset.

In addition, the role of transparency should be considered. YouTube does not disclose whether AI labels are applied by the platform or by content creators. This lack of disclosure makes it difficult to assess whether the higher number of labels on YouTube is due to proactive creator behavior or platform content moderation practices.

6.1.2. Impact of AI Labels on Engagement Metrics

As shown in Figure B.1, the engagement measures show similar distributions between labeled and unlabeled videos within each platform. In such cases, differences identified by the Mann-Whitney U test can be interpreted as differences in median values [17]. This complements the probability-based interpretation discussed in Section 5.3.

However, the potential relationship between AI labels and engagement should be interpreted with caution. Unlike TikTok, YouTube rarely displays AI labels directly in the video player. Within this dataset, only 0.8% of labeled videos included a visibly displayed label in the video player itself. Therefore, limited label visibility may reduce their impact on user behavior, as users may not notice the label.

Additionally, YouTube does not disclose whether a label was applied by the creator or the platform,

nor when it was added. Some videos may have received engagement before labeling, complicating causal interpretations. For TikTok, this issue was mitigated by analyzing only creator-applied labels, which cannot be removed or added after posting.

Despite these limitations, the statistical analysis reveals noteworthy patterns. The Mann-Whitney U test identified statistically significant differences in the like/view ratio, comment/view ratio, and view counts for both platforms. Only the share/view ratio (available exclusively for TikTok) showed no significant difference.

For both platforms, the median like/view ratio was lower for AI-labeled videos, suggesting that labeled content tends to receive fewer likes relative to views. The effect on the comment/view ratio differs between platforms: on YouTube Shorts non-labeled videos had a higher median, while on TikTok, labeled videos had a higher median. This may indicate that AI-labeled content on TikTok encourages more comments relative to views. This aligns with the research of [24], who attributed this effect to increased user curiosity toward labeled content. The share/view ratio showed no significant difference, indicating AI labels do not affect sharing behavior on TikTok. This finding is consistent with Li and Yang [50], who found no significant impact of AI labels on sharing intentions.

Finally, both platforms showed higher median view counts for labeled videos, but only on TikTok did the effect size exceed 0.1. This indicates a small but practically meaningful effect. In all other cases, effect sizes remained below 0.1. According to Cohen [16], this suggests negligible practical significance. A possible reason for the higher median view counts of labeled content is algorithmic influence. As discussed in Chapter 2, platform algorithms affect content visibility through moderation practices and recommendation systems. The finding that AI-labeled videos receive more views on both platforms suggests that algorithms may contribute to the increased visibility of such content. However, based on the current data, it remains unclear whether this results directly from algorithmic prioritization or from increased user interest in labeled content. Although only TikTok showed a practically significant effect, the higher view counts for labeled videos on both platforms may indicate some level of algorithmic prioritization.

Overall, it remains unclear to what extent the observed differences in engagement are directly caused by the presence of AI labels. Variations in label visibility and wording suggest that these effects may partly result from content characteristics rather than labeling alone.

6.1.3. Labeling Consistency

The labeling consistency analysis revealed clear differences between the two platforms. TikTok applies AI labels more accurately, with a precision of 98.10% and an accuracy of 75.00%. In contrast, YouTube Shorts shows much lower performance, with a precision of only 31.80% and an accuracy of 56.60%.

This lower performance for YouTube Shorts reflects frequent overlabeling, with a false positive rate of 34.09%. Of these cases, 99% fell under category Y1.5: Clearly Unrealistic Content. According to YouTube's policies [103], this type of content does not require a label. This overlabeling may result either from confusion among creators, who incorrectly label clearly unrealistic content, or from YouTube's automated detection systems applying labels regardless of the policy. In contrast, TikTok mandates labeling for this category and recorded only three false positives (0.97%), suggesting that YouTube creators may be unaware of the guideline not to label clearly unrealistic content.

In terms of underlabeling, YouTube Shorts showed a relatively low false negative rate (9.42%). This means that in most cases where a label was appropriate, it was applied. TikTok had a higher false negative rate (24.03%), primarily for category T5: Clearly Unrealistic Content. While underlabeling in this category may be less problematic, false negatives on both platforms also occurred in critical cases, such as the use of the likeness of real people, altered footage of real events, and realistic synthetic scenes.

Thematic coding further revealed that a part of the TikTok videos were tutorials or demonstrations explaining how to create AI-generated content. These videos were classified as instructional rather than AI-generated, which may partially explain the lower share of AI-generated content on TikTok, as shown in Figure 5.7.

It should be noted that the labeling consistency metrics presented here are based on thematic coding and the official platform guidelines. As such, they reflect an interpretation of what qualifies as a specific category of AI-generated content rather than an objective ground truth. Additionally, the analysis relies on a 5% sample of the dataset, which limits the generalizability of these findings.

6.2. Policy Implications and Platform Governance

The findings from this study raise important considerations for platform governance and EU policy initiatives. Despite existing transparency frameworks, such as the Digital Services Act (DSA) and the recently introduced EU AI Act, AI content labeling remains inconsistent and often lacks practical visibility.

As introduced in Chapter 2, the DSA is the primary policy for content moderation in the European Union [27]. This framework imposes stronger transparency requirements on very large online platforms (VLOPs) [46]. Another framework is the EU AI Act, which came into force in 2024 but does not place direct obligations on VLOPs. Instead, it requires AI content generators to include machine-readable labels in the visual content they produce [37]. As discussed in Section 2.3, both platforms use the C2PA standard to read these labels and detect AI-generated videos. However, there is a clear disconnect between the technical capabilities of C2PA and the platforms' guidelines for labeling AI-generated content (Appendix C). C2PA only detects the presence of embedded metadata indicating AI involvement but does not analyze the content itself for realism or potential to mislead. For example, a video created with an AI tool like ChatGPT's image generator might include a C2PA label regardless of whether it depicts a highly realistic human figure or a clearly fictional cartoon character.

Despite this technical detection capability, YouTube does not require creators to label clearly unrealistic content. Similarly, neither platform mandates labeling for the use of AI filters or visual effects. This results in a mismatch between what the technology can identify and what platform policies require creators to disclose. Such discrepancies can contribute to inconsistent labeling practices.

Moreover, the practical effectiveness of C2PA appears limited. Only 1.5% of TikTok videos in this study's dataset were labeled by the platform itself. This low prevalence suggests that many AI content generators may fail to apply machine-readable labels or that the C2PA method is not yet technically mature. Additionally, the widespread availability of tools to remove embedded metadata further undermines its effectiveness.

In conclusion, while the DSA and the EU AI Act represents a step towards increased transparency, the practical implementation of AI labeling appears to be inconsistent. Whether this is due to technical limitations of the C2PA-standard, non-compliance by AI content generators, or removal of metadata by creators remains unclear. However, the analysis indicates that platform-applied AI labels are infrequently used on TikTok. This is potentially also the case on YouTube, given the similar reliance on the C2PA standard. But since YouTube does not disclose whether labels were applied by the platform or the creator, this could not be verified.

6.3. Limitations

This research has several limitations that should be considered. First, the dataset was collected based on AI-related hashtags. This introduces a selection bias toward content explicitly tagged with these hashtags, which could have limited the diversity of the dataset. Creators wanting to avoid detection of AI-generated content may purposely avoid using these hashtags. In contrast, creators who are more transparent about their use of AI may be more likely to include AI-related hashtags and to label their content accordingly. This dynamic may have resulted in an over-representation of labeled videos.

Second, there may be a bias caused by platform algorithms that favor popular or recent content. As shown in the exploratory data analysis, the two platforms seem to differ in how content is displayed on the hashtag pages. YouTube showed an increase in videos around the data collection date. In contrast, TikTok showed a steeper curve which may suggest that TikTok prioritizes popular content more. These algorithmic preferences may have affected engagement metrics. Newer videos may have received less engagement due to limited exposure time, and popular videos may showed more engagement.

Third, the sample may have limited representativeness regarding demographic groups. AlAfnan [1] shows that social media behavior differs across various demographic and cultural contexts. Although this study used multiple VPN locations to collect data, these differences were not explicitly accounted for. Additionally, while both TikTok and YouTube operate on a global scale [21], this research primarily focused on the regulatory framework of the European Union. As a result, the dataset may include content that was produced for or primarily consumed in non-EU markets. This potential mismatch should be considered when interpreting the findings within the context of European policy frameworks.

Fourth, this study did not account for engagement behavior influenced by follower relationships. This may represent a relevant limitation, as prior research suggests that users tend to engage more

with content from creators they already follow [105]. This potentially places greater weight on social familiarity than on the presence of AI labels. Since this study did not distinguish between followers and non-followers when measuring engagement, this dynamic may have affected the observed results.

Moreover, the data only covers a specific time period from 01-11-2023 to 31-03-2025. This period was chosen because both platforms introduced AI labeling policies in November 2023 [30, 90]. However, it should be noted that several events within this time frame that may have influenced the outcomes of this research:

- According to TikTok's official platform documentation of C2PA was only implemented from May 2024 [92]. This may explain the under-representativeness of platform-applied AI labels on TikTok. However, at least one video posted in late 2023 was found with a platform applied AI label. This suggests that the automatic labeling system may have been applied retroactively. For YouTube, it is not officially documented when the automated AI labeling began.
- The EU AI Act came into force during the time period of the data and will be fully applicable only in 2026 [25]. As discussed, the AI Act requires AI content generators to apply machine-readable labels. This requirement, combined with the implementation of C2PA on both platforms, may have influenced the number of videos with platform-applied AI labels.

Lastly, the thesis did not control for the actual presence of AI-generated content in the engagement analysis. Instead, it only compared engagement between labeled and non-labeled videos. As thematic coding revealed, a proportion of the non-labeled videos were not AI-generated. This introduces a potential distorting factor, as differences in engagement may be driven by content type rather than the presence of absence of a label.

6.3.1. Ethical Considerations

This research has taken several measures to ensure that ethical considerations were fully integrated into the research process. There was a official application submitted for ethical approval to the Human Research Ethics Committee (HREC) at TU Delft and a Data Management Plan (DMP) was developed.

The study relied exclusively data from social media platforms that was publicly available. Although this data is public, some information could potentially lead to the re-identification of users or individuals appearing in videos. To mitigate this risk, no videos were downloaded or stored; only URLs and metadata were collected. However, since TikTok URLs contain usernames, the dataset is not publicly shared and is securely stored on the TU Delft Project Data Storage (U: drive). Access to the dataset was restricted to the research team.

Although scraping may violate the platform's terms of service, data collection was done responsibly and non-intrusively. Scraping methods were carefully applied by manually resolving captchas, rejecting cookie pop-ups and avoiding the use of headless browser applications. These measures attempted to minimize platform load. All data was collected solely for academic research purposes.

6.4. Recommendations

This research showed significant differences in how YouTube and TikTok apply AI labels. These differences include label visibility, source disclosure and application requirements. These findings led to the following recommendations for the platforms, policymakers, and future research.

6.4.1. Transparency and Platform-Applied AI Labels

YouTube currently does not distinguish between creator- and platform-applied AI labels. However, research shows that young adults tend to trust creator-applied labels less than those applied by platforms [80]. To enhance trust and transparency in terms of label source, YouTube should indicate whether a label was applied by the platform or the content creator.

In contrast, TikTok does make a distinction between creator- and platform-applied AI labels. However, this study found that only 1.5% of TikTok videos in the dataset had a platform-applied label. Since YouTube and TikTok both adopt the same method for applying platform labels, the assumption is that YouTube also has a really low rate of label applied by the platform. Therefore, both platforms would benefit from applying platform labels more frequently since they are seen as more credible.

If the C2PA standard does not detect sufficient AI-generated content, YouTube and TikTok could explore additional automated detection systems or invest in human moderation to improve labeling cov-

erage. Such efforts would demonstrate leadership in responsible AI governance and help the platforms stay ahead of evolving regulatory requirements.

Moreover, both platforms publicly state that their AI labeling initiatives aim to combat misinformation and promote responsible content creation [90, 101]. However, without transparent disclosure of label sources and a higher prevalence of platform-applied labels, these commitments risk being perceived as superficial. Increasing the visibility and credibility of AI labels is therefore not only beneficial for compliance but also essential to maintain user trust and protect the platforms' reputations in an increasingly AI-driven content landscape.

6.4.2. Standardization of Labeling Guidelines

Thematic coding revealed an inconsistent application of AI labels, particularly on YouTube. Nearly 35% of the YouTube sample was unnecessarily labeled, leading to a high rate of false positives. A likely cause is that YouTube does not mandate creators to label "Clearly Unrealistic Content" [101]. As a result, many videos receive a label even though they do not require one under YouTube's current guidelines.

In contrast, TikTok explicitly requires labeling for clearly unrealistic content and shows a significantly lower false positive rate (less than 1%). To reduce inconsistency and improve label accuracy, it is recommended that YouTube at least mandates labeling for clearly unrealistic content, aligning its policy with TikTok's clearer approach.

In addition, to promote uniformity between creator-applied and platform-applied labels, both platforms should consider extending their labeling requirements to all forms of AI-modified content. This includes:

- The use of AI-generated effects and filters.
- The addition of AI-generated elements to videos.

Currently, such modifications fall outside the labeling requirements of both platforms, even though they may be detectable through embedded metadata using the C2PA standard. However, whether AI-generated effects and filters consistently include such metadata depends on the specific tools used and their compliance with C2PA. By including these types of content modifications under their labeling policies, platforms can reduce user confusion and promote a more consistent application of AI labels.

6.4.3. Consideration for Emerging AI Content Generators

The EU AI Act currently mandates that AI content generation tools include machine-readable labels with generated content. However, this obligation does not have any application to the online platforms that distribute this content. Although the Digital Services Act (DSA) obliges platforms to take effective measures against misinformation [26], it does not explicitly mandate the disclosure of AI-generated content.

While AI-generated content does not necessarily qualify as misinformation, it is still important to make users aware when AI is involved in content creation. This need is reinforced by research indicating that distinguishing between human- and machine-generated content is becoming increasingly challenging [10, 14, 22, 29, 72].

It is therefore recommended that policymakers introduce a clear obligation for online platforms to disclose AI-generated content, either by extending the EU AI Act or by explicitly incorporating this requirement into the DSA.

In addition, given the current inconsistencies in how AI labels are displayed and described across platforms, policymakers should consider introducing clear guidelines on label placement and terminology. Existing research emphasizes that unclear or hidden labels can create an illusion of transparency and undermine the effectiveness of content moderation efforts [31]. Standardizing these practices would help ensure that AI labels are both visible and easily understood by users, regardless of the platform.

6.4.4. Recommendations for Further Research

Based on the findings and limitations of this study, several directions for future research are recommended:

- Investigate whether the higher median views for AI-labeled videos result from algorithmic prioriti-

zation in recommendation systems.

- Explore the contextual factors that may moderate the relationship between AI labels and user engagement. Given that prior studies report mixed results depending on content type and user motivations [24, 50], future research should investigate whether specific content categories influence how AI labels affect user behavior. This could help explain why this study found negligible engagement effects despite earlier conflicting findings in the literature.
- Include additional engagement metrics, such as watch time and video completion rates, to gain deeper insights into user behavior on short-form video platforms.
- Expand the dataset with a broader and more diverse range of hashtags, including non-AI-related topics, to assess whether AI-related hashtags influence labeling practices.
- Explore whether content creators strategically avoid AI labels. Prior research on content moderation avoidance through ”algospeak” [45] suggests similar strategies may emerge. Especially given the availability of tools that can remove embedded metadata.
- Assess how much AI-generated content contains misinformation and whether this type of content is frequently labeled, to better understand the effectiveness of AI labels in combating misinformation.
- Examine the practical implementation of the C2PA standard on short-form video platforms, including how and when C2PA metadata is embedded. Additionally, assess whether its implementation is consistent with the platforms' own labeling policies.

7

Conclusion

This study examined how AI labeling practices differ between TikTok and YouTube Shorts in terms of prevalence, consistency, and their influence on user engagement. The analysis was based on a dataset containing 13,163 videos, including 7,082 TikToks and 6,071 YouTube Shorts. Data was collected using a combination of scraping techniques and (un)official API access, based on AI-related hashtags. This thesis aimed to answer the following research question:

How do AI labeling practices on TikTok and YouTube Shorts differ in prevalence and consistency, and how do these labels impact user engagement?

The findings reveal that AI labels are more prevalent on YouTube Shorts than on TikTok. On TikTok, only 1.53% of videos carried a platform-applied AI label, while 35.34% had a creator-applied label. This suggests that the responsibility for AI labeling largely rests with content creators, with limited direct intervention from the platform itself. YouTube Shorts does not distinguish between the sources of AI labels, preventing users from knowing whether a label was applied by the platform or by the creator. Given that both platforms rely on the C2PA standard for detecting AI-generated content [92, 103], it is highly plausible that only a small fraction of labels on YouTube Shorts are platform-applied. Research indicates that platform-applied labels are generally perceived as more credible by users [80]. Therefore, the lack of source transparency and the low prevalence of platform-applied labels likely undermine the effectiveness of AI labeling practices on both platforms in supporting informed user judgments.

In terms of labeling consistency, thematic coding of 5% of the dataset revealed significant shortcomings on both platforms. AI labeling practices were applied inconsistently. YouTube Shorts exhibited a high level of overlabeling, likely caused by the absence of a requirement to label clearly unrealistic content. In contrast, TikTok showed higher levels of underlabeling, suggesting that AI-generated videos on this platform often remain unlabeled despite meeting the criteria for labeling. Furthermore, the labeling policies of both platforms contradict the capabilities of the C2PA standard they have adopted. C2PA can automatically detect AI-generated content through embedded metadata when such information is provided. It is therefore inconsistent that platforms do not require creators to label certain AI-generated content, even though the technology is capable of identifying it automatically.

The effect of AI labels on user engagement showed mixed results. Statistically significant differences in the median like/view ratio, comment/view ratio, and view counts were found on both platforms. However, only the effect on TikTok view counts showed a meaningful effect size. This effect was observed specifically for creator-applied labels, as these labels cannot be applied or removed after posting. On YouTube, the lack of transparency regarding label sources limits meaningful interpretation of the results. These findings suggest that the impact of AI labeling on user engagement is minimal and that only view counts on TikTok appear to be influenced, which may be explained by algorithmic prioritization.

If AI labels are not clearly visible or easily understood, they risk creating a false sense of transparency [31]. In addition, when labels are applied inconsistently and users are unaware of these inconsistencies, the implied truth effect may lead users to incorrectly perceive unlabeled content as more trustworthy [32, 67]. This is particularly problematic on TikTok, where underlabeling was observed,

increasing the likelihood that users will misinterpret unlabeled content as authentic. To maintain user trust and effectively moderate AI-generated content, platforms must prioritize improving the visibility, credibility, and consistency of their AI labeling practices.

Finally, this study underscores the need for ongoing policy efforts to establish clearer guidelines for AI content labeling and to strengthen platform accountability. Current European Commission policies primarily target AI content generation tools, while placing less emphasis on the platforms responsible for disseminating that content. However, this research shows that TikTok relies heavily on creator-applied labels, with limited oversight from the platform itself. It is likely that YouTube follows a similar approach, given its reliance on comparable labeling mechanisms. Future research should explore the drivers of user engagement with AI-labeled content, assess the role of algorithmic prioritization of AI labeled videos, and examine how technical standards like C2PA are applied in practice. Additionally, more research is needed on labeling effectiveness in combating misinformation and the strategies creators use to avoid AI labels.

These findings highlight an urgent need for more robust AI labeling policies. As generative AI becomes more integrated into online video content, platforms must not only define clear labeling guidelines, but also apply them consistently in practice. Currently, platform-applied labels are used rarely and creator-applied labels are used inconsistently, undermining their credibility and potentially confusing users. In addition, regulatory frameworks should be strengthened to ensure that AI-generated content is labeled accurately and visibly. Without such improvements, inconsistent labeling will continue to mislead users and erode trust in digital media.

References

[1] Mohammad Awad AlAfnan. “Cultural and Behavioral Insights into European Social Media Users: Platform Preferences and Personality Types”. In: *Studies in Media and Communication* 13.1 (Mar. 2025), pp. 17–30. ISSN: 2325808X. DOI: 10.11114/smc.v13i1.7306.

[2] S Altay and F Gilardi. “People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation”. In: *PNAS Nexus* 3.10 (2024). DOI: 10.1093/pnasnexus/pgae403. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85206300787&doi=10.1093%2fpnasnexus%2fpgae403&partnerID=40&md5=31825c581111120754dc9a3ca9cf8a27.

[3] Chittaranjan Andrade. *The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives Website: Quick Response Code*. Tech. rep. 2019.

[4] ASReview. *Join the movement towards fast, open, and transparent systematic reviews*. 2025. URL: https://asreview.nl/.

[5] Shubham Atreja, Libby Hemphill, and Paul Resnick. “Remove, Reduce, Inform: What Actions do People Want Social Media Platforms to Take on Potentially Misleading Content?” In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW2 (2023). DOI: 10.1145/3610082.

[6] K Balan et al. “EKILA: Synthetic Media Provenance and Attribution for Generative Art”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Vol. 2023-June. 2023, pp. 913–922. DOI: 10.1109/CVPRW59228.2023.00098.

[7] Monika Bickert. *Our Approach to Labeling AI-Generated Content and Manipulated Media*. Apr. 2024. URL: https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/.

[8] Andressa de Bittencourt Siqueira. “Automated detection of infringing content in Meta’s Oversight Board : How online content moderation is shaping new limits of freedom of expression”. In: *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI)*. Vol. 352. Gesellschaft fur Informatik (GI), 2024, pp. 183–197. ISBN: 9783885797463. DOI: 10.18420/inf2024{_}13.

[9] Beatriz Botero Arcila and Rachel Griffin. *Social media platforms and challenges for democracy, rule of law and fundamental rights*. Tech. rep. European Union, 2023.

[10] S D Bray, S D Johnson, and B Kleinberg. “Testing human ability to detect ’deepfake’ images of human faces”. In: *Journal of Cybersecurity* 9.1 (2023). DOI: 10.1093/cybsec/tyad011.

[11] H T Bui, V Filimonau, and H Sezerel. “AI-thenticity: Exploring the effect of perceived authenticity of AI-generated visual content on tourist patronage intentions”. In: *Journal of Destination Marketing and Management* 34 (2024). DOI: 10.1016/j.jdmm.2024.100956.

[12] Olivia Burrus, Amanda Curtis, and Laura Herman. “Unmasking AI: Informing Authenticity Decisions by Labeling AI-Generated Content”. In: *Interactions* 31.4 (Aug. 2024), pp. 38–42. DOI: 10.1145/3665321.

[13] John L. Campbell et al. “Coding In-depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement”. In: *Sociological Methods and Research* 42.3 (Aug. 2013), pp. 294–320. ISSN: 00491241. DOI: 10.1177/0049124113500475.

[14] B Chomanski and L Lauwaert. “Automated Propaganda: Labeling AI-Generated Political Content Should Not be Required by Law”. In: *Journal of Applied Philosophy* (2025). DOI: 10.1111/japp.70002.

[15] Coalition for Content Provenance and Authenticity. *Content Credentials C2PA Technical Specification*. Tech. rep. Sept. 2024. URL: https://c2pa.org/specifications/specifications/2.1/index.html.

[16] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 1988. ISBN: 9781134742707. DOI: 10.4324/9780203771587.

[17] Ronán M Conroy. *What hypotheses do "nonparametric" two-group tests actually test?* Tech. rep. 2. 2012, pp. 182–190.

[18] C L Cook, A Patel, and D Y Wohn. "Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms". In: *Frontiers in Human Dynamics* 3 (2021). DOI: 10.3389/fhumd.2021.626409. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85111662959&doi=10.3389%2ffhumd.2021.626409&partnerID=40&md5=beca77230ee64ae598630d065c32cd94.

[19] C Criddle and H. Murphy. *Meta debuts AI filmmaker in challenge to OpenAI's Sora*. Oct. 2024. URL: https://www.ft.com/content/00c9ce12-68f7-4fdd-a35f-8a66cc63420d?.

[20] D Delmonaco et al. ""What are you doing, TikTok?": How Marginalized Social Media Users Perceive, Theorize, and "Prove" Shadowbanning". In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW1 (2024). DOI: 10.1145/3637431. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85193285543&doi=10.1145%2f3637431&partnerID=40&md5=2caaf6a207ab501d0283e5ce4505e139.

[21] Stacey Jo Dixon. *Global social networks ranked by number of users 2024.* July 2024. URL: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.

[22] A Dmonte et al. "Classifying Human-Generated and AI-Generated Election Claims in Social Media". In: *Proceedings of the International Conference on Security and Cryptography*. 2024, pp. 237–248. DOI: 10.5220/0012797900003767.

[23] T Dobber et al. "The effect of traffic light veracity labels on perceptions of political advertising source and message credibility on social media". In: *Journal of Information Technology and Politics* 22.1 (2025), pp. 82–97. DOI: 10.1080/19331681.2023.2224316. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163150539&doi=10.1080%2f19331681.2023.2224316&partnerID=40&md5=4cdd1bb59b6d21407da143e0390148e5.

[24] D Du, Y Zhang, and J Ge. "Effect of AI Generated Content Advertising on Consumer Engagement". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 14039 LNCS. 2023, pp. 121–129. DOI: 10.1007/978-3-031-36049-7{_}9.

[25] European Comission. *AI Act*. Dec. 2024. URL: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[26] European Commission. *The Digital Services Act*. 2025. URL: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en.

[27] European Commission. *The impact of the Digital Services Act on digital platforms.* Apr. 2024. URL: https://digital-strategy.ec.europa.eu/en/policies/dsa-impact-platforms.

[28] Linda Farmus et al. "Effect size reporting and interpretation in social personality research". In: *Current Psychology* 42.18 (June 2023), pp. 15752–15762. ISSN: 19364733. DOI: 10.1007/s12144-021-02621-7.

[29] S A Fisher. "Something AI Should Tell You – The Case for Labelling Synthetic Content". In: *Journal of Applied Philosophy* 42.1 (2025), pp. 272–286. DOI: 10.1111/japp.12758.

[30] Jennifer Flannery O'Connor and Emily Moxley. *Our approach to responsible AI innovation*. Nov. 2023. URL: https://blog.youtube/inside-youtube/our-approach-to-responsible-ai-innovation/.

[31] Tobias Flattery and Christian B. Miller. ""Deepfakes and Dishonesty"". In: *Philosophy and Technology* 37.4 (Dec. 2024). ISSN: 22105441. DOI: 10.1007/s13347-024-00812-1.

[32] Melanie Freeze et al. "Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect". In: *Political Behavior* 43.4 (Dec. 2021), pp. 1433–1465. ISSN: 15736687. DOI: 10.1007/s11109-020-09597-3.

[33] Tarleton Gillespie. “Content moderation, AI, and the question of scale”. In: *Big Data and Society* 7.2 (July 2020). ISSN: 20539517. DOI: 10.1177/2053951720943234.

[34] Natasa Gisev, J. Simon Bell, and Timothy F. Chen. *Interrater agreement and interrater reliability: Key concepts, approaches, and applications*. May 2013. DOI: 10.1016/j.sapharm.2012.04.004.

[35] Google. *What is CAPTCHA?* 2025. URL: https://support.google.com/a/answer/1217728?hl=en.

[36] R Gorwa, R Binns, and C Katzenbach. “Algorithmic content moderation: Technical and political challenges in the automation of platform governance”. In: *Big Data and Society* 7.1 (2020). DOI: 10.1177/2053951719897945. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081602154&doi=10.1177%2f2053951719897945&partnerID=40&md5=55edb06da855fc32a840440f314c98be.

[37] Oskar J. Gstrein, Noman Haleem, and Andrej Zwitter. “General-purpose AI regulation and the European Union AI Act”. In: *Internet Policy Review* 13.3 (2024). ISSN: 21976775. DOI: 10.14763/2024.3.1790.

[38] Matt Halprin and Jennifer Flannery O’Connor. *On policy development at YouTube*. Dec. 2022. URL: https://blog.youtube/inside-youtube/policy-development-at-youtube/.

[39] C Harris et al. “”Honestly, I Think TikTok has a Vendetta Against Black Creators”: Understanding Black Content Creator Experiences on TikTok”. In: *Proceedings of the ACM on Human-Computer Interaction* 7.CSCW2 (2023). DOI: 10.1145/3610169. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174501340&doi=10.1145%2f3610169&partnerID=40&md5=f67de7fc917bc0852db94c1e36eb4dd5.

[40] Traci Hong et al. “Effects of #coronavirus content moderation on misinformation and anti-Asian hate on Instagram”. In: *New Media and Society* (Feb. 2023). ISSN: 14617315. DOI: 10.1177/14614448231187529.

[41] Natalya Izotova, Mariia Polishchuk, and Kateryna Taranik-Tkachuk. “Discourse analysis and digital technologies: (TikTok, hashtags, Instagram, YouTube): universal and specific aspects in international practice”. In: *Revista Amazonia Investiga* 10.44 (Sept. 2021), pp. 198–206. DOI: 10.34069/ai/2021.44.08.19.

[42] Shagun Jhaver et al. “Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter”. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021). ISSN: 25730142. DOI: 10.1145/3479525.

[43] S Jiang, R E Robertson, and C Wilson. “Reasoning about political bias in content moderation”. In: *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*. 2020, pp. 13669–13672. DOI: 10.1609/aaai.v34i09.7117. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85086566176&doi=10.1609%2faaai.v34i09.7117&partnerID=40&md5=d9e95d5666d793491a377464594eabeb.

[44] Amelia Johns et al. “Labelling, shadow bans and community resistance: did meta’s strategy to suppress rather than remove COVID misinformation and conspiracy theory on Facebook slow the spread?” In: *Media International Australia* (2024). ISSN: 1329878X. DOI: 10.1177/1329878X241236984.

[45] D Klug, E Steen, and K Yurechko. “How Algorithm Awareness Impacts Algospeak Use on TikTok”. In: *ACM Web Conference 2023 - Companion of the World Wide Web Conference, WWW 2023*. 2023, pp. 234–237. DOI: 10.1145/3543873.3587355. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159584154&doi=10.1145%2f3543873.3587355&partnerID=40&md5=781397b1f752f807db778aae96313be1.

[46] L Kosters and O J Gstrein. “TikTok and Transparency Obligations in the EU Digital Services Act (DSA) – A Scoping Review”. In: *Zeitschrift fur Europarechtliche Studien* 27.1 (2024), pp. 110–145. DOI: 10.5771/1435-439X-2024-1-110. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85188104799&doi=10.5771%2f1435-439X-2024-1-110&partnerID=40&md5=a6c7d992f73d57df13b44ec4f28f9276.

[47] Deepak Kumar et al. “Designing Toxic Content Classification for a Diversity of Perspectives”. In: USENIX Association, Aug. 2021. ISBN: 9781939133250. URL: https://www.usenix.org/conference/soups2021/presentation/kumar.

[48] J Richard Landis and Gary G Koch. *The Measurement of Observer Agreement for Categorical Data*. Tech. rep. 1. 1977, pp. 159–174. URL: https://about.jstor.org/terms.

[49] Donghee N. Lee et al. *Impact of Financial Disclosures and Health Warnings on Youth and Young Adult Perceptions of Pro-E-cigarette Instagram Posts*. Feb. 2024. DOI: 10.1093/ntr/ntad219.

[50] F Li and Y Yang. “Impact of Artificial Intelligence-Generated Content Labels On Perceived Accuracy, Message Credibility, and Sharing Intentions for Misinformation:Web-Based, Randomized, Controlled Experiment”. In: *JMIR Formative Research* 8 (2024). DOI: 10.2196/60024.

[51] Chen Ling, Krishna P Gummadi, and Savvas Zannettou. “”Learn the Facts about COVID-19”: Analyzing the Use of Warning Labels on TikTok Videos”. In: *Proceedings of the International AAAI Conference on Web and Social Media* 17.1 (2023). DOI: 10.1609/icwsm.v17i1.22168.

[52] Y Liu, S Wang, and G Yu. “The nudging effect of AIGC labeling on users’ perceptions of automated news: evidence from EEG”. In: *Frontiers in Psychology* 14 (2023). DOI: 10.3389/fpsyg.2023.1277829.

[53] H Ma, W Huang, and A R Dennis. “Unintended Consequences of Disclosing Recommendations by Artificial Intelligence versus Humans on True and Fake News Believability and Engagement”. In: *Journal of Management Information Systems* 41.3 (2024), pp. 616–644. DOI: 10.1080/07421222.2024.2376381.

[54] N Marchal and H Au. “”Coronavirus EXPLAINED”: YouTube, COVID-19, and the Socio-Technical Mediation of Expertise”. In: *Social Media and Society* 6.3 (2020). DOI: 10.1177/2056305120948158. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089392620&doi=10.1177%2f2056305120948158&partnerID=40&md5=79cb2a56ac52b47ee15233077f09adab.

[55] Arash Marioriyad and Pouria Ramazi. “Optimizing Accuracy, Recall, Specificity, and Precision Using ILP”. In: *Mathematics* 13.7 (Apr. 2025). ISSN: 22277390. DOI: 10.3390/math13071059.

[56] Pamela San Martín. “Meta’s Oversight Board: Challenges of Content Moderation on the Internet”. In: *Erasmus Law Review* 2023.2 (2023), pp. 124–139. ISSN: 22102671. DOI: 10.5553/ELR.000253.

[57] Mary L McHugh. “Interrater reliability: the kappa statistic.” In: *Biochemia medica* 22.3 (2012), pp. 276–82. ISSN: 1330-0962.

[58] Liv McMahon, Zoe Kleinman, and Courtney Subramanian. *Facebook and Instagram get rid of fact checkers*. Jan. 2025. URL: https://www.bbc.com/news/articles/cly74mpy8klo.

[59] Mendeley. *Mendeley Reference Manager*. 2025. URL: https://www.mendeley.com/reference-management/reference-manager.

[60] J Mo et al. “Towards Trustworthy Digital Media In The Aigc Era: An Introduction To The Upcoming IsoJpegTrust Standard”. In: *IEEE Communications Standards Magazine* 7.4 (2023), pp. 2–5. DOI: 10.1109/MCOMSTD.2023.10353009.

[61] Scott Monteith et al. “Artificial intelligence and increasing misinformation”. In: *British Journal of Psychiatry* 224.2 (Feb. 2024), pp. 33–35. ISSN: 14721465. DOI: 10.1192/bjp.2023.136.

[62] J Nassetta and K Gross. “State media warning labels can counteract the effects of foreign disinformation”. In: *Harvard Kennedy School Misinformation Review* 1.7 (2020). DOI: 10.37016/mr-2020-45. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110672413&doi=10.37016%2fmr-2020-45&partnerID=40&md5=52c018b06c8f3bf1f068775ca5cdc306.

[63] Hellina Hailu Nigatu and Inioluwa Deborah Raji. “”I Searched for a Religious Song in Amharic and Got Sexual Content Instead”: Investigating Online Harm in Low-Resourced Languages on YouTube.” In: *2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024*. Association for Computing Machinery, Inc, June 2024, pp. 141–160. ISBN: 9798400704505. DOI: 10.1145/3630106.3658546.

[64] Olga Papadopoulou, Evangelia Kartsounidou, and Symeon Papadopoulos. “COVID-Related Misinformation Migration to BitChute and Odysee”. In: *Future Internet* 14.12 (Dec. 2022). ISSN: 19995903. DOI: 10.3390/fi14120350.

[65] Pat Pataranutaporn et al. “AI-generated characters for supporting personalized learning and well-being”. In: *Nature Machine Intelligence* 3.12 (Dec. 2021), pp. 1013–1022. ISSN: 2522-5839. DOI: 10.1038/s42256-021-00417-9.

[66] Nikolas Pautz, Benita Olivier, and Faans Steyn. “The use of nonparametric effect sizes in single study musculoskeletal physiotherapy research: A practical primer”. In: *Physical Therapy in Sport* 33 (Sept. 2018), pp. 117–124. ISSN: 18731600. DOI: 10.1016/j.ptsp.2018.07.009.

[67] Gordon Pennycook et al. “Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention”. In: *Psychological Science* 31.7 (July 2020), pp. 770–780. ISSN: 14679280. DOI: 10.1177/0956797620939054.

[68] Joe Phua and Dongjae Lim. “Can warning labels mitigate effects of advertising message claims in celebrity-endorsed Instagram-based electronic cigarette advertisements? Influence on social media users’ E-cigarette attitudes and behavioral intentions”. In: *Journal of Marketing Communications* 29.5 (2023), pp. 455–475. ISSN: 14664445. DOI: 10.1080/13527266.2022.2037008.

[69] Monroe E Price and Joshua M Price. “Building Legitimacy in the Absence of the State: Reflections on the Facebook Oversight Board”. In: *International Journal of Communication* 17 (2023), pp. 3315–3325. URL: http://ijoc.org..

[70] Kevin Proudfoot. “Inductive/Deductive Hybrid Thematic Analysis in Mixed Methods Research”. In: *Journal of Mixed Methods Research* 17.3 (July 2023), pp. 308–326. ISSN: 15586901. DOI: 10.1177/15586898221126816.

[71] Robert Rosenthal. “Meta-analytic procedures for social science research”. In: *Educational Researcher* 15.8 (Oct. 1986), pp. 18–20. ISSN: 0013-189X. DOI: 10.3102/0013189X015008018.

[72] S Rossi et al. “Are Deep Learning-Generated Social Media Profiles Indistinguishable from Real Profiles?” In: *Proceedings of the Annual Hawaii International Conference on System Sciences*. Vol. 2023-January. 2023, pp. 134–143.

[73] Ian Russell. *Debate: More, not less social media content moderation? How to better protect youth mental health online*. Sept. 2024. DOI: 10.1111/camh.12717.

[74] Rose Marie Santini, Débora Salles, and Bruno Mattos. “Recommending instead of taking down: YouTube hyperpartisan content promotion amid the Brazilian general elections”. In: *Policy and Internet* 15.4 (Dec. 2023), pp. 512–527. ISSN: 19442866. DOI: 10.1002/poi3.380.

[75] F Saurwein and C Spencer Smith. “Automated trouble: The role of algorithmic selection in harms on social media platforms”. In: *Media and Communication* 9.4 (2021), pp. 222–233. DOI: 10.17645/mac.v9i4.4062. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119197995&doi=10.17645%2fmac.v9i4.4062&partnerID=40&md5=4179e72d4d5da059403c5e5fd98ef1ba.

[76] A K Saxena. “Quantitative Measurement of Bias in AI-Generated Content: A Comprehensive Narrative Literature Review”. In: *International Symposium on Technology and Society, Proceedings*. 2024. DOI: 10.1109/ISTAS61960.2024.10732696. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85210232780&doi=10.1109%2fISTAS61960.2024.10732696&partnerID=40&md5=64956938a6eca9b4a47279b760bad4b3.

[77] J Seering. “Reconsidering Community Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2 (2020). DOI: 10.1145/3415178. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85094199961&doi=10.1145%2f3415178&partnerID=40&md5=915d18e496311d8a8d6e406f899a9457.

[78] F Sharevski et al. “Talking Abortion (Mis)information with ChatGPT on TikTok”. In: *Proceedings - 8th IEEE European Symposium on Security and Privacy Workshops, Euro S and PW 2023*. 2023, pp. 594–608. DOI: 10.1109/EuroSPW59978.2023.00071. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85168236254&doi=10.1109%2fEuroSPW59978.2023.00071&partnerID=40&md5=87d0e2ddb73b83f5c321f41386a7774d.

[79] Mohamed R. Shoaib et al. “Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models”. In: *ICCA 2023 - 2023 5th International Conference on Computer and Applications, Proceedings*. Institute of Electrical and Electronics Engineers Inc., 2023. ISBN: 9798350303254. DOI: 10.1109/ICCA59364.2023.10401723.

[80] E Shusas. “Designing Better Credibility Indicators: Understanding How Emerging Adults Assess Source Credibility of Misinformation Identification and Labeling”. In: *DIS 2024 - Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 2024, pp. 41–44. DOI: 10.1145/3656156.3665126. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85198901204&doi=10.1145%2f3656156.3665126&partnerID=40&md5=5f7d9d95d36239285c245eb36249f8f5.

[81] Nathan A Silver et al. “Improving Enforcement Measures and Establishing Clear Criteria: A Content Analysis of Tobacco-Brand-Owned Instagram Accounts”. In: *Nicotine and Tobacco Research* 26.9 (Aug. 2024), pp. 1175–1182. ISSN: 1469-994X. DOI: 10.1093/ntr/ntae052.

[82] Ben Steel and Alexei Abrahams. *networkdynamics/pytok: Initial working version of library*. July 2024. DOI: 10.5281/zenodo.12802714. URL: https://doi.org/10.5281/zenodo.12802714.

[83] E Steen, K Yurechko, and D Klug. “You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok”. In: *Social Media and Society* 9.3 (2023). DOI: 10.1177/20563051231194586. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85169663074&doi=10.1177%2f20563051231194586&partnerID=40&md5=b0c29106e4f6648b773102dbae7c7e4c.

[84] Gail M. Sullivan and Richard Feinn. “ Using Effect Size—or Why the P Value Is Not Enough ”. In: *Journal of Graduate Medical Education* 4.3 (Sept. 2012), pp. 279–282. ISSN: 1949-8349. DOI: 10.4300/jgme-d-12-00156.1.

[85] Kah Yee Tai, Jasbir Dhaliwal, and Vinod Balasubramaniam. “Leveraging Mann–Whitney U test on large-scale genetic variation data for analysing malaria genetic markers”. In: *Malaria Journal* 21.1 (Dec. 2022). ISSN: 14752875. DOI: 10.1186/s12936-022-04104-x.

[86] David Teather. *TikTokAPI (Version 7.0.0) [Computer software]*. 2025. URL: https://github.com/davidteather/tiktok-api.

[87] F Temmermans and L Rosenthol. “Adopting the JPEG universal metadata box format for media authenticity annotations”. In: *Proceedings of SPIE - The International Society for Optical Engineering*. Vol. 11842. 2021. DOI: 10.1117/12.2597651.

[88] TikTok. *#ai*. 2025. URL: https://www.tiktok.com/tag/ai.

[89] TikTok. *Community Guidelines*. Apr. 2024. URL: https://www.tiktok.com/community-guidelines/en/overview.

[90] TikTok. *New labels for disclosing AI-generated content*. Sept. 2023. URL: https://newsroom.tiktok.com/en-us/new-labels-for-disclosing-ai-generated-content.

[91] TikTok. *Our approach to content moderation*. 2025. URL: https://www.tiktok.com/transparency/en/content-moderation/.

[92] TikTok. *Partnering with our industry to advance AI transparency and literacy*. May 2024. URL: https://newsroom.tiktok.com/en-us/partnering-with-our-industry-to-advance-ai-transparency-and-literacy.

[93] TikTok Help Center. *About AI-generated content*. 2025. URL: https://support.tiktok.com/en/using-tiktok/creating-videos/ai-generated-content.

[94] C. Vallance. *Meta is ditching fact checkers for X-style community notes. Will they work?* Jan. 2025. URL: https://www.bbc.com/news/articles/c4g93nvrdz7o.

[95] Evelyn Wan. *Laboring in Electronic and Digital Waste Infrastructures: Colonial Temporalities of Violence in Asia*. Tech. rep. 2021, pp. 2631–2651. URL: http://ijoc.org..

[96] Y Wei and G Tyson. “Understanding the Impact of AI-Generated Content on Social Media: The Pixiv Case”. In: *MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, pp. 6813–6822. DOI: 10.1145/3664647.3680631. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85209792753&doi=10.1145%2f3664647.3680631&partnerID=40&md5=3e7bd7e062af4a4e0b78c7202f941270.

[97] Chloe Wittenberg et al. “Labeling AI-Generated Content: Promises, Perils, and Future Directions”. In: *An MIT Exploration of Generative AI* (2024). DOI: 10.21428/e4baedd9.0319e3a6.

[98] D Wong and L Floridi. “Meta’s Oversight Board: A Review and Critical Assessment”. In: *Minds and Machines* 33.2 (2023), pp. 261–284. DOI: 10.1007/s11023-022-09613-x. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140612129&doi=10.1007%2fs11023-022-09613-x&partnerID=40&md5=9b5a3d8afea7a5f71823b735a16cedd4.

[99] Scott Wright. “Government-run online discussion fora: Moderation, censorship and the shadow of control”. In: *British Journal of Politics and International Relations* 8.4 (Nov. 2006), pp. 550–568. ISSN: 13691481. DOI: 10.1111/j.1467-856X.2006.00247.x.

[100] YouTube. *#ai*. 2025. URL: https://www.youtube.com/hashtag/ai/shorts.

[101] YouTube. *How we’re helping creators disclose altered or synthetic content*. Mar. 2024. URL: https://blog.youtube/news-and-events/disclosing-ai-generated-content/.

[102] YouTube. *YouTube Data API*. 2025. URL: https://developers.google.com/youtube/v3.

[103] YouTube Help. *Disclosing use of altered or synthetic content*. 2025. URL: https://support.google.com/youtube/answer/14328491#zippy=%2Cexamples-of-content-that-creators-dont-have-to-disclose%2Cexamples-of-content-that-creators-need-to-disclose.

[104] YouTube Help. *YouTube’s Community Guidelines*. 2025. URL: https://support.google.com/youtube/answer/9288567?hl=en.

[105] Savvas Zannettou et al. “Analyzing User Engagement with TikTok’s Short Format Video Recommendations using Data Donations”. In: *Conference on Human Factors in Computing Systems - Proceedings*. New York: Association for Computing Machinery, May 2024. ISBN: 9798400703300. DOI: 10.1145/3613904.3642433.

[106] Cindy C. Zhang et al. *Debate: Social media content moderation may do more harm than good for youth mental health*. Feb. 2024. DOI: 10.1111/camh.12689.

A

Search Queries/Methods

Table A.1: Search Queries Used in Literature Review

Sub-Question	Search Query/Method	Criteria	Identified	Relevant	Used
(1) Content Moderation	`TITLE-ABS-KEY (("content moderation" OR "misinformation label" OR "moderation techniques" OR "warning label" OR "fact-check label") AND ("short-video platform" OR "TikTok" OR "Meta" OR "YouTube"))`	>2020 English	n = 140	n = 51	n = 33
	Backward Snowballing	>2020			n = 6
	Grey Literature	Only official websites			n = 5

(2) AI labels	`TITLE-ABS-KEY (("AI-generated" OR "AIGC" OR "synthetic media" OR "AI") AND label* AND ("TikTok" OR "YouTube" OR "social media" OR "short-form video platforms"))`
	
	`OR TITLE-ABS-KEY ("C2PA" OR "Coalition for Content Provenance and Authenticity")`
	
	`OR TITLE-ABS-KEY (label* AND ("AI-generated" OR "AIGC" OR "synthetic media") AND (engagement OR prevalence OR perception OR trust))`
		>2020	n = 193	n = 24	n = 12

	Backward Snowballing	>2020			n = 10
	Grey Literature	Only official websites or reports			n = 8

B

Data Collection and Processing Details

B.1. Final Dataset Structure

Table B.1: Column Structure of the Final Dataset

Column Name	Present for YouTube	Present for TikTok
`url`	X	X
`ai_label`	X	
`views`	X	X
`likes`	X	X
`comments`	X	X
`shares`		X
`saved`		X
`hashtags`	X	X
`platform`	X	X
`publishedAt`	X	X
`sensitive_topic`	X	
`publishedAt_date`	X	X
`publishedAt_day`	X	X
`publishedAt_month`	X	X
`like_view_ratio`	X	X
`comment_view_ratio`	X	X

B.2. Distribution of Engagement Metrics

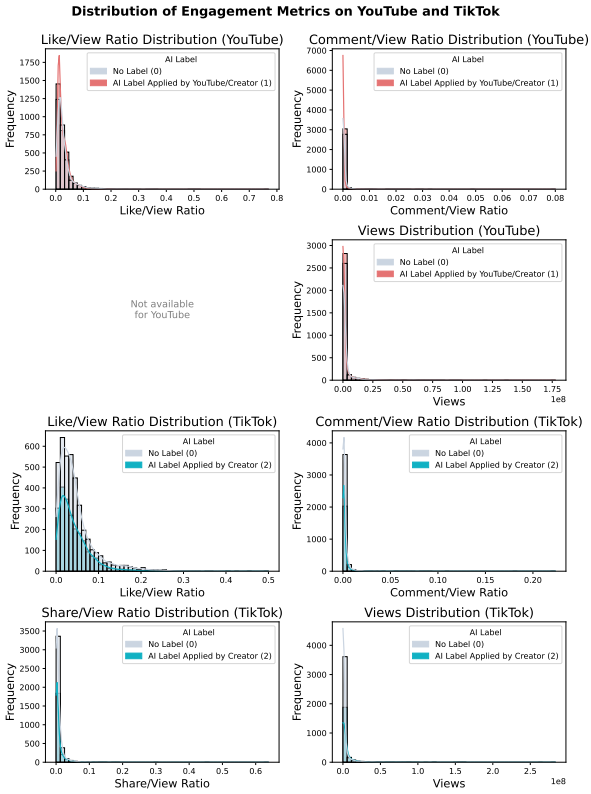


Figure B.1: Distribution of Engagement Metrics on YouTube Shorts and TikTok

B.3. Hashtag Selection

Search Hashtag	Top 18 Co-occurring Hashtags	Count
#ai	#ai	1133
	#shorts	376
	#fyp	193
	#aiart	182
	#trending	167
	#cute	152
	#cat	125
	#funny	125
	#aigenerated	124
	#viral	124
	#squidgame	108
	#squidgame2	97
	#aivideo	97
	#animals	80
	#artificialintelligence	75
	#love	70
	#marvel	70
	#spiderman	69
#aiart	#aiart	982
	#ai	739
	#cat	432
	#cute	429
	#fyp	315
	#shorts	278
	#catsoftiktok	264
	#viral	258
	#tiktok	219
	#catlover	210
	#funny	198
	#aigenerated	154
	#poorcat	150
	#aivideo	142
	#cats	140
	#aicat	136
	#kitten	136
	#cutecat	128
#aigenerated	#ai	757
	#aigenerated	727
	#marvel	654
	#avengers	645
	#spiderman	555
	#shorts	542
	#trending	362
	#viral	341
	#dc	264
	#superheroes	260
	#aiart	239
	#dog	204
	#aivideo	180
	#fyp	165
	#cartoonfinalamazing	138
	#puppy	137
	#doglover	135
	#midjourney	110

Table B.2: Top 18 hashtags for #ai, #aiart, and #aigenerated (1-3 of 25)

Used Hashtag	Top 18 Co-occurring Hashtags	Count
#aicat	#cat	1140
	#cute	861
	#aicat	651
	#ai	625
	#catsoftiktok	578
	#viral	527
	#tiktok	452
	#fyp	430
	#shorts	406
	#kitten	394
	#cutecat	380
	#catlover	378
	#funny	350
	#cats	332
	#aiart	299
	#trend	227
	#unitedstates	224
	#india	220
#aivideo	#ai	650
	#aivideo	598
	#aiart	210
	#fyp	174
	#aigenerated	174
	#automobile	152
	#shorts	142
	#aiedits	108
	#aivfx	97
	#trending	86
	#automotiveedit	83
	#artificialintelligence	80
	#funny	78
	#viral	70
	#aivideos [a]	66
	#cat	64
	#hornsound	64
	#viralvideo	61
#aiedtis	#ai	716
	#aiedits	564
	#aivideo	454
	#aivfx	416
	#automobile	409
	#automotiveedit	375
	#aivideos	115
	#fyp	112
	#aiart	67
	#edit	63
	#transformers	59
	#aigenerated	57
	#viral	54
	#foryou	45
	#aitools	38
	#artificialintelligence	35
	#monster	34
	#foryoupage	32

Table B.3: Top 18 hashtags for #aicat, #aivideo, and #aiedtis (4-6 of 25)

[a] Too much alike as #aivideo

Used Hashtag	Top 18 Co-occurring Hashtags	Count
#artificialintelligence	#artificialintelligence	718
	#ai	530
	#trendingshorts	440
	#babyfashion	221
	#ytshorts	220
	#aibaby	219
	#trendybaby	219
	#babystyle	218
	#cutebabyclothes	216
	#aifashionmodel	212
	#viralbaby	211
	#aifashionshow[a]	208
	#aibabayfashionshow[b]	204
	#shorts	203
	#adorableoutfits	199
	#babywardrobe	198
	#babyclothingtrends	196
	#babyfashioninspo	195
#aitools	#aitools	788
	#ai	648
	#chatgpt	192
	#shorts	177
	#artificialintelligence	166
	#aiart	147
	#youtubeshorts	102
	#trending	91
	#aivideo	90
	#film	87
	#aigeneratedfilm	86
	#coding	58
	#viral	58
	#productivity	54
	#aistorytelling	54
	#aitool[c]	53
	#college	51
	#content	49
#aivfx	#ai	806
	#aiedits	585
	#aivideo	578
	#aivfx	546
	#automobile	486
	#automotiveedit	439
	#vfx	196
	#aiart	128
	#aivideos	116
	#transformers	103
	#shorts	81
	#fyp	70
	#aitools	67
	#artificialintelligence	60
	#aigenerated	58
	#monster	56
	#vfxvideo	56
	#vfxmagicvideo	56

Table B.4: Top 18 hashtags for #artificialintelligence, #aitools, and #aivfx (7-9 of 25)

Used Hashtag	Top 18 Co-occurring Hashtags	Count
#aibaby	#aibaby	842
	#trendingshorts	830
	#babyfashion	444
	#aifashionmodel	427
	#ytshorts	409
	#trendybaby	402
	#viralbaby	402
	#aifashionshow	401
	#ai	400
	#cutebabyclothes	398
	#babystyle	389
	#babywardrobe	382
	#stylishbabies	382
	#babyfashiontips	379
	#babyoutfits	379
	#fashiontipsforparents	378
	#littlefashionista	378
	#fashionforbabies	378
#aifashionmodel	#trendingshorts	764
	#ai	508
	#aifashionmodel	406
	#ytshorts	381
	#aifashionshow	380
	#fashiontipsforparents	363
	#littlefashionista	362
	#aiedits	340
	#adorableoutfits	331
	#artificialintelligence	313
	#aibaby	311
	#babyfashion	310
	#cutebabyclothes	306
	#trendybaby	306
	#babystyle	305
	#viralbaby	298
	#babywardrobe	295
	#stylishbabies	292
#aigeneratedfilm	#aigeneratedfilm	840
	#film	743
	#aiart	602
	#aitools	413
	#ai	389
	#aistorytelling	379
	#shortfilm	377
	#scifi	375
	#sciencefiction	366
	#future	364
	#cinematography	362
	#content	360
	#filmproduction	360
	#digitalfrontier	359
	#scifishortfilm	357
	#experimentalfilm	357
	#aivideo	152
	#aigenerated	138

Table B.5: Top 18 hashtags for #aibaby, #aifashionmodel, and #aigeneratedfilm (10-12 of 25)

[a] Too much alike as #aifashionmodel
[b] Too much alike as #aifashionmodel
[c] Too much alike as #aitools

Used Hashtag	Top 18 Co-occurring Hashtags	Count
#aistorytelling	#aistorytelling	448
	#ai	445
	#aiart	266
	#shorts	226
	#artificialintelligence	164
	#aistory	146
	#fyp	142
	#creativeai	138
	#aigenerated	136
	#aicontent	115
	#aiinnovation	113
	#chatgpt	112
	#aianimation	98
	#youtubeshorts	90
	#aishorts[a]	87
	#digitalcreativity	84
	#aivideo	82
	#aiproductions	77
#creativeai	#ai	591
	#aiart	488
	#creativeai	472
	#shorts	289
	#artificialintelligence	272
	#digitalart	188
	#chatgpt	187
	#aistorytelling	185
	#aigenerated	163
	#aivideo	160
	#aicontent	156
	#aianimation	153
	#aiinnovation	141
	#fyp	140
	#openai	130
	#sora	125
	#digitalcreativity	116
	#gpt	113
#aiinnovation	#ai	540
	#aiinnovation	440
	#artificialintelligence	434
	#shorts	326
	#halloween	199
	#trending	199
	#aiart	184
	#aitrends	177
	#shortsfeed	176
	#innovation	159
	#aigenerated	159
	#machinelearning	156
	#aistorytelling	153
	#airevolution	149
	#aicommunity	148
	#aicontent	147
	#creativeai	143
	#chatgpt	137

Table B.6: Top 18 hashtags for #aistorytelling, #creativeai, and #aiinnovation (13-15 of 25)

[a]YouTube-specific: only on YouTube it is called 'shorts'

Used Hashtag	Top 18 Co-occurring Hashtags	Count
#aicontent	#aicontent	505
	#ai	478
	#shorts	361
	#artificialintelligence	251
	#aiart	248
	#aigenerated	204
	#aivideo	167
	#aiinnovation	133
	#aistorytelling	131
	#chatgpt	124
	#viralvideo	118
	#creativeai	116
	#fyp	105
	#openai	100
	#aishorts	100
	#aianimation	98
	#digitalcreativity	95
	#youtubeshorts	91
#aiproductions	#artificialintelligence	492
	#chatgpt	481
	#creativeai	425
	#openai	420
	#aistorytelling	414
	#aicontent	412
	#sora	405
	#aiinnovation	402
	#aiproductions	401
	#digitalcreativity	397
	#gpt	396
	#movieprompts	396
	#promptedmovies	396
	#sorachannel	396
	#soragineratedvideos	396
	#faz3	350
	#fazza	350
	#ai	326
#aianimation	#aianimation	566
	#ai	410
	#shorts	282
	#aiart	275
	#cat	153
	#fyp	126
	#aigenerated	110
	#catlover	108
	#funnycats	104
	#aivideo	103
	#cats	100
	#catvideos	92
	#funny	92
	#cuteanimals	90
	#animation	88
	#petlovers	73
	#viralcats	72
	#apt	72

Table B.7: Top 18 hashtags for #aicontent, #aiproductions, and #aianimation (16-18 of 25)

Used Hashtag	Top 18 Co-occurring Hashtags	Count
#aitrends	#ai	726
	#aitrends	547
	#shorts	341
	#trending	307
	#artificialintelligence	246
	#shortsfeed	235
	#halloween	232
	#aigenerated	224
	#aiart	221
	#aivideo	181
	#aicommunity	178
	#viral	175
	#fyp	175
	#aishorts	167
	#aiinnovation	163
	#youtubeshorts	151
	#machinelearning	147
	#deeplearning	134
#aicommunity	#ai	713
	#aicommunity	576
	#aiart	546
	#shorts	277
	#midjourney	258
	#artificialintelligence	244
	#halloween	235
	#trending	208
	#aigenerated	201
	#aitrends	181
	#shortsfeed	177
	#aivideo	176
	#aiartcommunity	164
	#fyp	147
	#capcut	143
	#beauty	133
	#machinelearning	131
	#shorteo	124
#aistory	#aistory	771
	#ai	599
	#shorts	449
	#hen	219
	#cat	162
	#youtubeshorts	139
	#fyp	133
	#story	126
	#catstory	119
	#animals	112
	#aiart	110
	#trending	109
	#aivideo	107
	#aicat	102
	#chicken	101
	#pigeon	94
	#cartoon	91
	#viral	90

Table B.8: Top 18 hashtags for #aitrends, #aicommunity, and #aistory (19-21 of 25)

Used Hashtag	Top 18 Co-occurring Hashtags	Count
#airevolution	#airevolution	874
	#ai	685
	#artificialintelligence	387
	#shorts	349
	#machinelearning	268
	#trending	223
	#youtubeshorts	214
	#futuretech	194
	#techinnovation	186
	#aiinnovation	183
	#halloween	173
	#deeplearning	165
	#futureofai	160
	#viralvideo	158
	#shortsfeed	152
	#aicommunity	145
	#aitrends	135
	#viral	134
#aiartcommunity	#aiart	833
	#ai	793
	#aiartcommunity	732
	#aiartwork	533
	#aiartist	488
	#cat	481
	#cute	325
	#catlovers	294
	#pets	263
	#catart	253
	#catshorts	251
	#funny	244
	#aigenerated	242
	#bing	236
	#digitalart	216
	#illustration	204
	#love	198
	#graphicdesign	194
#aiartwork	#ai	785
	#aiart	771
	#aiartwork	628
	#cat	442
	#aiartcommunity	438
	#aiartist	417
	#cute	300
	#funny	286
	#catlovers	262
	#pets	247
	#catart	229
	#catshorts	206
	#aigenerated	205
	#bing	198
	#love	182
	#shorts	176
	#illustration	170
	#digitalart	163

Table B.9: Top 18 hashtags for #airevolution, #aiartcommunity, and #aiartwork (22-24 of 25)

Used Hashtag	Top 18 Co-occurring Hashtags	Count
#aiartwork	#ai	959
	#cat	678
	#aiart	643
	#aiartist	619
	#aiartwork	534
	#cute	476
	#aiartcommunity	471
	#catlovers	412
	#pets	379
	#catart	367
	#funny	349
	#catshorts	343
	#bing	305
	#illustration	279
	#love	267
	#graphicdesign	259
	#aicat	252
	#shorts	228

Table B.10: Top 18 hashtags for #aiartwork (25 of 25)

C

Codebooks

C.1. Categories – YouTube

Y1) AI-generated Content

Y1.1 Using the likeness of a realistic person *[Label]*

- Digitally generating or altering content to replace the face of one individual with another's[1]
- Simulating audio to make it sound as if a medical professional gave advice when the professional did not actually give that advice[1]
- Cloning someone else's voice for voiceovers or dubs[1]
- Digitally altering audio to make it sound as if a popular singer missed a note in their live performance[1]
- Depicting a public figure stealing something that they did not steal or admitting to stealing something when they did not make that admission[1]

Y1.2 Altering footage of real events or places *[Label]*

- Such as making it appear as if a real building caught fire, or altering a real cityscape to make it appear different than in reality [2]
- Synthetically generating extra footage of a real place, like a video of a surfer in Maui for a promotional travel video[1]

Y1.3 Generating realistic scenes *[Label]*

- Digitally altering a famous car chase scene to include a celebrity who wasn't in the original film[1]
- Synthetically generating extra footage of a real place, like a video of a surfer in Maui for a promotional travel video[1]
- Showing a realistic depiction of fictional major events, like a tornado moving toward a real town[2]
- Making it look like a real person has been arrested or imprisoned[1]
- Showing a realistic depiction of a missile fired towards a real city[1]
- Making it appear as if hospital workers turned away sick or wounded patients[1]

Y1.4 Generating music *[Label]*

- Synthetically generating music[1]

Y1.5 Clearly unrealistic content

- Animation or someone riding a unicorn through a fantastical world[2]
- Green screen used to depict someone floating in space[1]

[1] YouTube Help [103]
[2] YouTube [101]

Y1.6 Effects / filters

- Using effects to enhance previously recorded audio[1]
- Color adjustment or lighting filters[2]
- Special effects like background blur or vintage effects[2]
- Beauty filters or other visual enhancements[2]
- Video sharpening, upscaling or repair, and voice or audio repair[1]

Y1.7 Adding elements to a video

- Synthetically generating or extending a backdrop to simulate a moving car[1]

Y1.8 Improve videos or AI help

- Production assistance, like using generative AI tools to create or improve a video outline, script, thumbnail, title or infographic[1]
- Caption creation[1]
- Idea generation[1]

Y1.9 Synthetic self-likeness

- Cloning one's own voice to create voiceovers or dubs[1]

Y2) Sensitive Topic

Y2.1 Discusses sensitive topics

- Elections, conflicts, public health crises, or officials[3]

C.2. Categories – TikTok

T1) AI-generated Content

T1.1 Using the likeness of a realistic person[4] *[Label]*

- Video featuring a real person speaking, whose image, voice, and/or words are altered or modified by AI[5]
- The primary subjects are portrayed doing something they didn't do, for example, dancing[5]
- The primary subjects are portrayed saying something they didn't say, for example, AI-generated speech[5]
- The appearance of the primary subject(s) has been substantially altered, such that the original subject(s) is no longer recognizable, for example, with an AI face-swap[5]

T1.2 Altering footage of real events or places *[Label]*

- Video or image featuring a scene or event that occurred in the real world, but has been altered or modified by AI[5]

T1.3 Generating realistic scenes *[Label]*
T1.4 Generating Music *[Label]*
T1.5 Clearly unrealistic content *[Label]*

- Entirely AI-generated videos or images of real or fictional people, places, and events[5]

T1.6 Minor retouching[5]
T1.7 Effects/filters[6]

[3] Flannery O'Connor and Moxley [30]
[4] TikTok Help Center [93]
[5] TikTok [89]

D

Engagement Hypotheses

The following null (H_0) and alternative (H_1) hypotheses were tested using Mann-Whitney U tests. Each hypothesis examines whether there is a statistically significant difference of engagement metrics between videos with and without an AI label within a specific platform.

YouTube

- H_0: There is no difference in the like/view ratio between labeled and non-labeled.
  H_1: There is a difference in the like/view ratio between labeled and non-labeled YouTube videos.
- H_0: There is no difference in the comment/view ratio between labeled and non-labeled YouTube videos.
  H_1: There is a difference in the comment/view ratio between labeled and non-labeled YouTube videos.
- H_0: There is no difference in the number of views between labeled and non-labeled YouTube videos.
  H_1: There is a difference in the number of views between labeled and non-labeled YouTube videos.

TikTok

- H_0: There is no difference in the like/view ratio between labeled and non-labeled TikTok videos.
  H_1: There is a difference in the like/view ratio between labeled and non-labeled TikTok videos.
- H_0: There is no difference in the comment/view ratio between labeled and non-labeled TikTok videos.
  H_1: There is a difference in the comment/view ratio between labeled and non-labeled TikTok videos.
- H_0: There is no difference in the share/view ratio between labeled and non-labeled TikTok videos.
  H_1: There is a difference in the share/view ratio between labeled and non-labeled TikTok videos.
- H_0: There is no difference in the number of views between labeled and non-labeled TikTok videos.
  H_1: There is a difference in the number of views between labeled and non-labeled TikTok videos.