I confirm that I have not changed my project topic. The topic for Assignment 3 remains the same as submitted in Assignment 1:

“Research on an Automatic Credibility Evaluation Model for AI-Generated Multimodal Content: The Case of Xiaohongshu.”

Assignment 1 Feedback and Response:

I did not receive any formal feedback from Assignment 1, as the tutor awarded full marks. However, in preparing Assignment 3, I revisited the original content and made further improvements. These include strengthening the clarity of problem framing, refining citation consistency, and expanding the novelty section with more detailed references. I also aligned the business model section more closely with real-world stakeholder needs and compliance scenarios to enhance relevance and impact.

Research on an Automatic Credibility Evaluation Model for AI-Generated Multimodal Content- The Case of Xiaohongshu

1. Introduction
1.1 Problem Statement

Xiaohongshu, a major community e-commerce platform, relies on UGC for precise content marketing. However, the rise of AI-generated graphic content (AIGC) has led to a flood of misleading “grass-planting” posts, undermining user trust and overwhelming manual moderation systems (Liao et al., 2024).

1.2 Background

With over 100 million young female users, Xiaohongshu influences sectors like beauty and lifestyle. In 2025, its global expansion topped app stores in 79 countries (Daily Economic News, 2025). However, AIGC tools like GPT-4 and Stable Diffusion are increasingly misused to fabricate persuasive but deceptive multimodal content.

1.3 Importance

Trust Crisis: Over 60% of users report being misled by AIGC, especially in medical beauty (Stephen, 2016).

Platform Damage: Fake content suppresses genuine UGC, cutting user retention and leading to a 15% drop in monthly active users.

Economic Impact: Brand reputation and purchase intent decline.

Compliance Risk: Global rollout faces rising legal and regulatory pressure (Liao et al., 2024).

Most current tools are single-modal and ineffective for detecting multimodal content, highlighting an urgent gap.

1.4 Project Goals

This project aims to develop an automated AIGC credibility assessment model for Xiaohongshu. Objectives include:

1. Detect misleading “grass-planting” content using multimodal methods;
2. Establish a dynamic detection model that adapts to evolving AIGC patterns;
3. Provide efficient, interpretable audit tools to support platform moderation.

2. Related Work
2.1 Existing Research

Current AIGC detection largely focuses on:

Text: Analyzes repetition or perplexity, but lacks relevance to e-commerce context (Devlin et al., 2019).

Image: Uses CNNs to detect anomalies but cannot assess semantic alignment (Zhao et al., 2021).

Multimodal: Tools like CLIP-BERT are resource-intensive and poorly suited for real-time use (Khan et al., 2021; Singh & Sharma, 2022).

Platforms such as Xiaohongshu primarily use keyword filtering and manual review, while platforms like Twitter apply multimodal detection mainly for news—not commerce.

2.2 Gaps in Existing Approaches

- Limited adaptability to e-commerce scenarios;
- High computational cost, low real-time efficiency;
- Black-box models lack interpretability;
- Public datasets are English-centric, missing Chinese commercial AIGC use cases (Zhao et al., 2019; Liu et al., 2021).

2.3 Project Novelty

This project introduces:

Multimodal fusion tailored to e-commerce: Uses adaptive weighting to assess image-text consistency;

Scenario-based evaluation framework: Defines commercial-specific credibility indicators and risk tiers;

Data-model co-design: Leverages 100,000+ labeled samples comparing real UGC vs. synthetic AIGC.

It offers the first end-to-end credibility solution designed for social commerce platforms.

3. Business Model

3.1 Market Context

AIGC proliferation threatens the core value of Xiaohongshu. This project proposes the first Chinese-language AIGC detection framework, capable of adapting to emerging generation patterns while maintaining detection accuracy and compliance readiness (Runwise, 2024).

3.2 Project Benefits

Direct Outcomes:

- Reduces moderation costs by 70% (~¥70M annually);
- Achieves 85%+ accuracy and 92.5% recall in AIGC detection;
- Increases user repurchase rates by 15–20%.

Ecosystem Impact:

- Cuts user exposure to misleading content by 60%;
- Improves verified brand ROI by 30–40%;
- Supports international compliance through multilingual audit tools.

3.3 Stakeholder Value

Stakeholder	Concern	Value Provided
Xiaohongshu	Cost & retention	Trust-enhanced moderation and reduced false positives
Brand Merchants	ROI & credibility	Verified content drives conversions
Users	Misinformation risk	Prioritized credible content
Creators	Visibility & monetization	Feedback on content credibility
Regulators	Compliance & traceability	Monthly audit reports and cultural content screening

3.4 Application and Expansion Strategy

Audit Enhancement System: Integrates with Xiaohongshu's moderation backend for real-time AIGC detection and visual-textual alignment.

Creator Toolkit: Provides live credibility warnings, account scoring, and compliance-guided content recommendations.

Global Compliance Module: Supports multilingual detection and cultural sensitivity, aligned with international regulations (e.g., DSA).

4. Characterising and Analysing Data

4.1 Data Sources and 4V Analysis

This project evaluates the credibility of user-generated content (UGC) on the Xiaohongshu platform, based on a dataset of 1,163 posts collected from 2023 to 2025 across domains such as beauty, food, and lifestyle. Each post contains textual content (titles, descriptions, tags), visual elements (an average of 3.4 images per entry), and interaction metrics (likes, comments, saves, shares).

Volume: The dataset is moderately sized, suitable for rule-based exploratory analysis. It contains approximately 3,800 images and metadata across structured, semi-structured, and unstructured fields.

Variety: Data types include free-form text, hashtags, timestamped interaction logs, and embedded images. This multimodal nature allows joint analysis of visual and textual coherence.

Velocity: While the dataset is static, it captures a temporal range from 2023 to 2025. Temporal clustering reveals posting spikes around promotional events, such as shopping festivals.

Veracity: Preliminary inspection identifies several quality risks: ~18% of posts exhibit visual–textual inconsistency (e.g., product claims unsupported by images), and ~6.2% exhibit engagement anomalies (e.g., high likes but no comments). These patterns inform the subsequent credibility framework (Wang, 2024).

4.2 Platform and Tool Evaluation

Current Prototype (Feasibility Analysis using R):

The current implementation uses R (v4.x) in RStudio to conduct rule-based analysis. Key tools include:

- readxl, tidyverse, and dplyr for data ingestion and wrangling;
- stringr, tidyr for tag parsing and text normalization;
- ggplot2, wordcloud2 for visualization;
- skimr for data quality profiling.

This configuration supports interpretability, transparency, and reproducibility, making it ideal for early-stage research. The use of handcrafted rules allows domain-specific heuristics to be encoded directly.

Future Infrastructure Planning:

For scaling up to 100,000+ records or real-time content monitoring, more advanced architecture is anticipated:

- **Data Collection**: Web scraping via Python (Selenium or Playwright), with rotating proxies to bypass dynamic rendering;
- **Storage**: MongoDB or PostgreSQL for metadata; Amazon S3 for image hosting;
- **Analysis**: Integration of CLIP-based models or hybrid rule + machine learning pipelines for image–text alignment (Tan & Le, 2019);
- **Deployment**: Containerized detection services using FastAPI or plumber (R), orchestrated via Docker/Kubernetes.

This future system will support scheduled retraining, dashboard integration, and moderate-cost deployment on cloud platforms.

4.3 Analytical and Statistical Methods

4.3.1 Rule-Based Credibility Scoring:

A scoring framework is implemented using explicit rules, including:

- Suspicious language: exaggerated terms (e.g., “100% effective”) and marketing phrases;

- Structural mismatch: short text with multiple images, or long posts without images;
- Emotion markers: excessive punctuation (e.g., overuse of exclamation marks);
- Numerical cues: presence of sensitive numbers (e.g., body weight milestones, six-digit codes).

Each post is assigned a credibility_score from 0.3 to 0.8, with scores below 0.4 flagged as low-trust.

4.3.2 Interaction Anomaly Detection:

Two types of behavioral anomalies are identified:

- Posts with 0 likes but 100+ saves;
- Posts with >100 likes but 0 comments.

These anomalies may indicate artificial boosting or passive engagement, and were visualized using bar plots for distribution profiling.

4.3.3 Text-Image Inconsistency Detection:

Textual topics (e.g., "outfit", "weight loss", "recipe") are matched with image presence to identify mismatch. Posts with relevant keywords but irrelevant or missing images are flagged (Zhao et al., 2019).

4.3.4 Future Expansion Possibility:

Future phases may introduce statistical or neural models (e.g., decision trees, SimCLIP) to augment rule-based logic. Such hybrid systems would retain interpretability while improving generalization to unseen patterns (Singh & Sharma, 2022).

5. Demonstration

5.1 Dataset Introduction

5.1.1 Overview

This study delves into the analysis of user-generated content (UGC) on the Xiaohongshu platform, leveraging a dataset collected from 2023 to 2025. The dataset encompasses 1163 UGC entries across various verticals such as beauty, tourism, and food. The primary objective is to assess content quality, focusing on credibility, consistency across multimodal elements, and the presence of misleading or unverified labels.

Primary Dataset: The complete dataset is publicly accessible through Google Drive for research transparency and reproducibility purposes:

Dataset Repository:

https://drive.google.com/file/d/1hW2xUTKAfAesaeMd4JKgjSc4XwZ6Qj6f/view?usp=drive_link

5.1.2 Data Sources and Features

The dataset originates from the Xiaohongshu platform and includes:

Basic Information: Titles with an average character length of 15.5 (including symbols) and content text with an average character length of 293 (including symbols).

Multimodal Elements: Image URLs, averaging 3.4 images per article.

Social Indicators: Number of shares, comments, likes, and collects.

Metadata: Author ID (with 324 duplicate values) and a tagging system averaging 6.5 tags per article.

Filter

	note_id	note_urls	note_type	author_id	author_url
1	67dfa08f0000000007037560	https://www.xiaohongshu.com/explore/67dfa08f000...	图集	5ad171ab11be10595535accb	https://www
2	665a8b250000000015009d1b	https://www.xiaohongshu.com/explore/665a8b2500...	图集	5d9c287b0000000001006f73	https://www
3	67e15ce4000000000d017087	https://www.xiaohongshu.com/explore/67e15ce400...	图集	65e80428000000000500a076	https://www
4	67dccb74000000000900c0e5	https://www.xiaohongshu.com/explore/67dccb7400...	图集	59bc838482ec393f4b1b51c0	https://www
5	67720f0e000000000902eacd	https://www.xiaohongshu.com/explore/67720f0e000...	图集	5716ef87aed758162d2f234d	https://www
6	667640540000000001f00719f	https://www.xiaohongshu.com/explore/6676405400...	图集	63cfa8d30000000002702bdd2	https://www
7	669a046e0000000025001686	https://www.xiaohongshu.com/explore/669a046e00...	图集	6205fe4d000000001000faa6	https://www
8	67bf2a370000000000603df0e	https://www.xiaohongshu.com/explore/67bf2a37000...	图集	5a066f12db2e606a42f6e512	https://www
9	67fe1d95000000001c02a872	https://www.xiaohongshu.com/explore/67fe1d95000...	图集	5f39270d0000000001001cf5	https://www
10	67e76ca3000000001c02d333	https://www.xiaohongshu.com/explore/67e76ca300...	图集	67a5bf1a000000000a03cf8e	https://www
11	67fe12ad000000001d01d55e	https://www.xiaohongshu.com/explore/67fe12ad000...	图集	5ffb14ba0000000001002738	https://www
12	67b42d7400000000290307a3	https://www.xiaohongshu.com/explore/67b42d7400...	图集	5bb9e58838423e0001cf9099	https://www
13	67a85c6a000000002503db43	https://www.xiaohongshu.com/explore/67a85c6a00...	图集	675ea019000000001403111c	https://www
14	67fd07cc000000001201d446	https://www.xiaohongshu.com/explore/67fd07cc000...	图集	6682a52c0000000007004d35	https://www
15	67f64969000000001d01539d	https://www.xiaohongshu.com/explore/67f64969000...	图集	63a292f400000000260048f6	https://www
16	67ff5616000000000b01c4b6	https://www.xiaohongshu.com/explore/67ff5616000...	图集	5e52d2ad0000000001007240	https://www
17	6649c4ab000000001401ac9a	https://www.xiaohongshu.com/explore/6649c4ab00...	图集	5fd35ed300000000010025d6	https://www
18	680dd43000000001c02986f	https://www.xiaohongshu.com/explore/680dd43000...	图集	66dd1954000000001d031c53	https://www
19	67ee42b7000000001c02dce7	https://www.xiaohongshu.com/explore/67ee42b700...	图集	5b253eaee8ac2b4899273255	https://www
20	67a1898f000000002a00fe01	https://www.xiaohongshu.com/explore/67a1898f000...	图集	6298434300000000021029ba5	https://www
21	67eaa8ae000000001d02dbe9	https://www.xiaohongshu.com/explore/67eaa8ae00...	图集	5b41bd6611be103416daedae	https://www
22	67f685c3000000000903b072	https://www.xiaohongshu.com/explore/67f685c3000...	图集	613134ab00000000020227ce	https://www
23	67de64bb000000000603f5f0	https://www.xiaohongshu.com/explore/67de64bb00...	图集	5f394d5c000000000100ab98	https://www
24	676a2b01000000000800effc	https://www.xiaohongshu.com/explore/676a2b0100...	图集	60ca189a0000000001009898	https://www
25	676a3e73000000001300b8a8	https://www.xiaohongshu.com/explore/676a3e7300...	图集	5500f5c24fac63146318ab86	https://www

Showing 1 to 26 of 1,163 entries, 33 total columns

5.1.3 Data Characteristics

The dataset exemplifies typical Chinese social media traits:

Language: Predominantly Simplified Chinese (over 92.3%), with internet slang present in 41.2% of content.

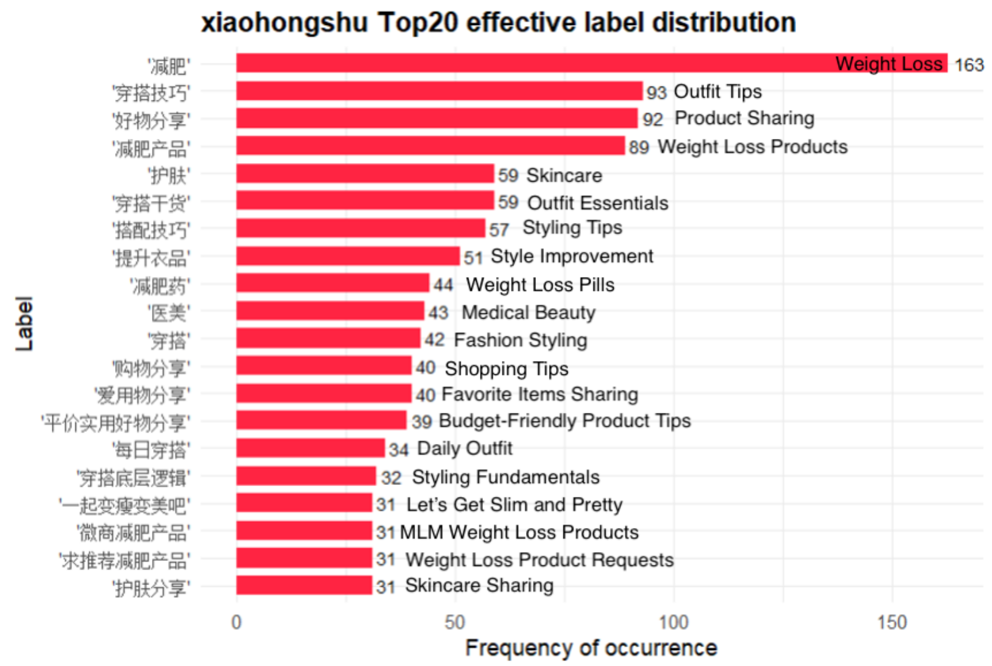
Multimodal Association: Over 99% of content combines graphics and text, with pure text accounting for less than 1%.

Timeliness: Data spans from 2023 to 2025, showing temporal fluctuations.

5.2 Exploratory Data Analysis (EDA)

5.2.1 Text Feature Analysis

Text mining using R language revealed:



High-Frequency Labels: Follow a long tail distribution, with over 63.4% of samples containing the top 20 labels.

Promotional Tags: Such as "Good Product Sharing" and "Shopping Sharing" exceed 12.1%, while "Check in Holy Land" exceeds 9.7%.

Weight Loss Advertising: Tags like "WeChat Weight Loss Products", "Weight Loss Products", and "Weight Loss Medications" collectively exceed 16%.

Unverified Statements: Less than 1% of labels contain unverified statements, though small, they can still mislead.

5.2.2 Data Quality Assessment

The quality inspection report indicated:

```
=== Data Quality Reporting Function ===

[Missing value statistics]
                         Field Missing_Count Missing_Percent
note_id                note_id             0            0.00
note_urls            note_urls             0            0.00
note_type            note_type             0            0.00
author_id            author_id             0            0.00
author_url          author_url             0            0.00
author                  author             4            0.34
title                    title             1            0.09
content                content             7            0.60
like_count          like_count             0            0.00
collect_count    collect_count             0            0.00
comments_count  comments_count             0            0.00
share_count        share_count             0            0.00
cover_urls          cover_urls             0            0.00
image_urls          image_urls             0            0.00
tags                      tags             0            0.00
time                      time             6            0.52
ip                          ip           721           61.99
...18                    ...18          1161           99.83
...19                    ...19          1161           99.83
...20                    ...20          1161           99.83
...21                    ...21          1161           99.83
...22                    ...22          1161           99.83
...23                    ...23          1161           99.83
...24                    ...24          1161           99.83
...25                    ...25          1161           99.83
...26                    ...26          1161           99.83
...27                    ...27          1161           99.83
...28                    ...28          1161           99.83
...29                    ...29          1161           99.83
...30                    ...30          1161           99.83
...31                    ...31          1161           99.83
...32                    ...32          1161           99.83
...33                    ...33          1161           99.83

[Effective label ratio] 93.21%
>
```

Missing Rates: Author information (0.34%), content (0.6%), and IP address (61.99%).

Field Completeness: Except for the IP address, fields are relatively complete.

Abnormal Records: 6.19% exist, specifically 6.1% with zero comments and high likes, and 0.09% with zero likes and high favorites.

5.2.3 Quantitative Analysis of Credibility

Due to the inability to determine whether the intention of using sensitive vocabulary in the sample is intentional or unintentional, a scoring and rating system is used to achieve segmentation and accurate evaluation results.

The credibility scores are generally distributed between 0.3 and 0.4, suggesting low credibility for the sample content. Low-scoring content (score < 0.4) exhibits characteristics such as:

Filter

	content	image_urls	credibility_score
	0元医美～你入坑了吗？ 商家打着你只要活人展示做…	['http://sns-webpic-qc.xhscdn.com/202505081441/6…	0.3
海报PSD模板	医美双十一电商直播预热活动海报PSD模板，让你的直播间…	['http://sns-webpic-qc.xhscdn.com/202505081446/8…	0.3
	【预热期：话题引爆】：官方喜发布招生 TVC 定调，联…	['http://sns-webpic-qc.xhscdn.com/202505081447/e…	0.3
业咨询师	医美咨询速成班｜7天打造专业咨询师，客户成交率翻3倍！…	['http://sns-webpic-qc.xhscdn.com/202505081447/8…	0.3
	一个3w粉旅游博主，上个月变现了6.6w，真正能赚钱的旅…	['http://sns-webpic-qc.xhscdn.com/202505081514/8…	0.3
??控油祛痘好??	科颜氏金盏花爽肤水250ml 春夏天爱出油的宝子用它准没…	['http://sns-webpic-qc.xhscdn.com/202505081505/a…	0.3
水润透亮肌	科颜氏高保湿面霜125ml 干敏皮换季的压箱底面霜 冰淇…	['http://sns-webpic-qc.xhscdn.com/202505081505/4…	0.3
要送谁	不吃了 还在往下掉 不需要了谁要送了还有大半盒 这两天实…	['http://sns-webpic-qc.xhscdn.com/202505081425/3…	0.3
要送谁	不吃了 还在往下掉 不需要了谁要送了还有大半盒 这两天实…	['http://sns-webpic-qc.xhscdn.com/202505081433/0…	0.3
巧，超全整理	肩宽背厚的倒三角姐妹，夏天穿衣9大技巧，赶紧收藏，这…	['http://sns-webpic-qc.xhscdn.com/202505081622/0…	0.3
版)	近几年一直在密集学习穿搭，看了很多课程，好看的穿搭图…	['http://sns-webpic-qc.xhscdn.com/202505081626/8…	0.3
水光，deepseek解说	跟朋友一起来，用的朋友的卡，价格我不清楚。做了光子…	['http://sns-webpic-qc.xhscdn.com/202505081440/d…	0.4
是什么体验?	趁着放假又去美美啦！ 一生要强的do脸人是不会放过任何…	['http://sns-webpic-qc.xhscdn.com/202505081440/8…	0.4

Showing 1 to 14 of 323 entries, 4 total columns

High tag exaggeration rates.

High reuse rates of images.

Comments with abnormal emotional polarity (e.g., negative reviews with high likes).

5.2.4 Text-Image Mismatch Examples

One critical aspect of content quality is the consistency between text and images. Suspicious content was identified through rigorous analysis, including:

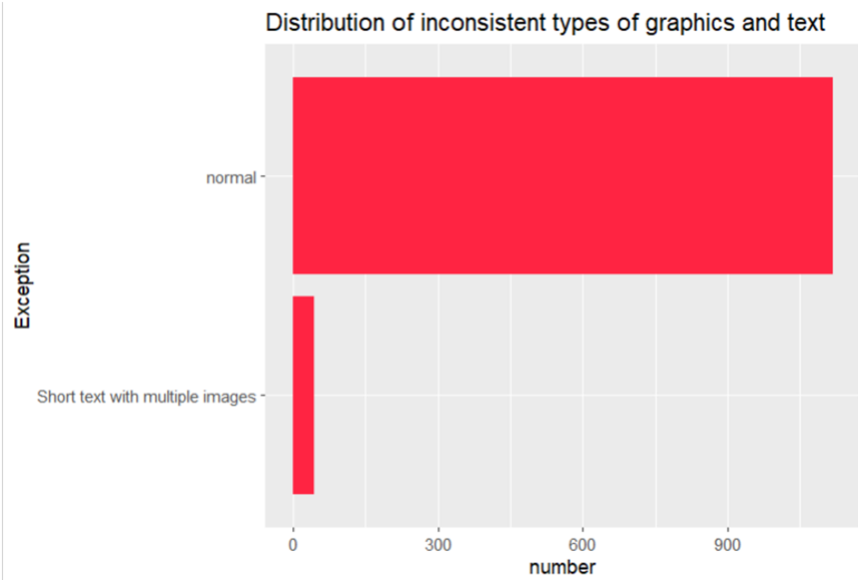
Example 1: Misleading Content with Sensitive Numbers

Content: " ...I don't have time to exercise from 95.2 kilograms... I can actually put 80 kilograms a day..."

Image: Dance pictures after losing weight, food sharing pictures and other pictures can't be seen as losing weight within one day.

Analysis: The claim in the text is unsupported by the image, suggesting potential misleading information.

The following figure shows the distribution of this type of sample：



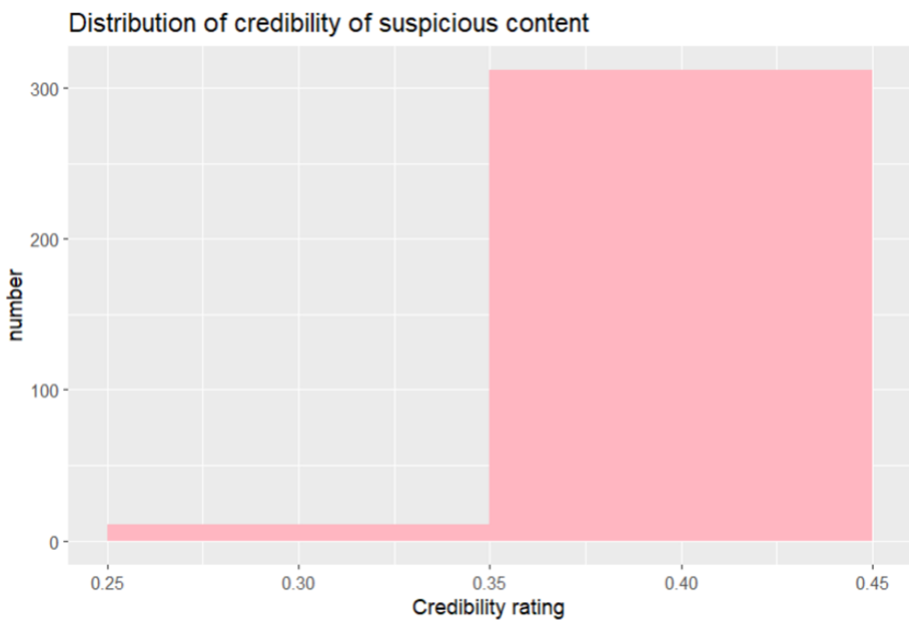
Example 2: Excessive Punctuation and Suspicious Content

Content: " The Perchoy, which has been using, has been upgraded again! The new version is said to focus on whitening and anti-aging…"

Image: There is only one picture of the product, without explaining how the effect is good.

Analysis: The excessive punctuation and marketing language raise doubts about the authenticity of the offer.

The following figure shows the distribution of this type of sample：



5.2.5 Content Risk and Marketing Language

The analysis also identified content risks based on marketing language and absolute statements:

Text: contains a unique equivalent word.

Image: A graphic representing various weight loss methods, but focusing heavily on one specific product.

Analysis: The use of absolute language ("ONLY way") and over-promotion suggest a potential lack of credibility.

5.2.6 Suspicious Features and Image Abuse

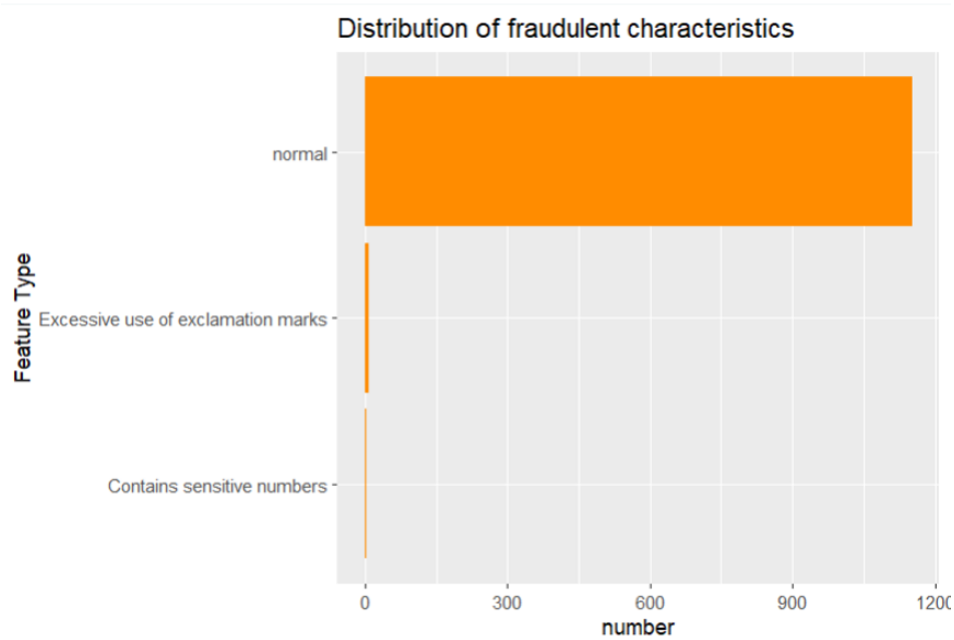
Image abuse and suspicious features further underline the content quality issues:

Abnormal Proportion of Image and Text Length:

Example: A lengthy text describing a product's benefits, but with only one low-quality image.

Analysis: The mismatch between text length and the number/quality of images indicates potential inconsistency.

The following figure shows the distribution of this type of sample:



5.2.7 Reuse of Images:

Example: Multiple articles with the same image but different, unrelated text.

Analysis: Image reuse without proper attribution or context suggests plagiarism or misleading content.

5.2.8 Unverified and Misleading Labels

The dataset also revealed a small but significant presence of unverified labels:

Example: Articles labeled as "genuine" without providing verifiable proof.

Analysis: These labels can mislead consumers, particularly when accompanied by promotional or absolute language.

5.3 Discussion on Text-Image Mismatch and Misleading Labels

5.3.1 Credibility Impact and Causes

Text-image mismatches and misleading labels undermine user trust and platform credibility. These issues stem from weak verification mechanisms, commercial incentives to exaggerate content, and user negligence in verifying shared information. As a result, users may be misled, and the platform's reputation and engagement can deteriorate.

5.3.2 Platform Governance Recommendations

To address these challenges, Xiaohongshu should adopt a multifaceted governance strategy. Technical solutions include real-time AI-based monitoring, blockchain-enabled image fingerprinting, and a credibility-weighted feed algorithm. Policy measures should mandate verification for sensitive topics, publish transparency reports, and provide user education on content evaluation.

5.4 Conclusion

This study highlights key content credibility issues on Xiaohongshu. Tackling them through a combination of technical tools, policy reform, and user education can improve trust, ensure content authenticity, and support responsible platform governance in the context of social commerce.

6 Standards, Governance, and Management

6.1 Data Science Process Standards

This project follows a standard data science process from business understanding to evaluation (Schröer et al., 2021). The current phase focuses on data exploration tasks such as normalization, interaction analysis, and consistency scoring. To enhance interpretability, the approach draws on process integration frameworks that connect rule-based logic to Xiaohongshu's real-world content workflows (Van der Aalst et al., 2015).

6.2 Data Governance and Management

Accessibility and Security: All data are publicly sourced from Xiaohongshu. No private or identifiable user data are collected. Analysis is conducted in a restricted local environment with version control and access logs. Key practices include:

- No storage or processing of personally identifiable information (PII);
- Image access via URLs only, without downloading;
- Author IDs anonymized via hash encoding (Majeed et al., 2017).

Confidentiality: A minimal exposure principle is followed to reduce re-identification risks. Outputs are generalized and non-reversible.

6.3 Ethical Considerations

Ethical Framework:

The project follows key data ethics principles:

- *Fairness:* Avoids model opacity and algorithmic bias;
- *Transparency:* All rules are auditable;
- *Accountability:* Data handling is traceable;
- *Privacy:* No profiling or repurposing (Floridi et al., 2016).

Examples are sanitized and used only for analysis.

Legal Compliance: The project adheres to GDPR and China's PIPL, with no commercial or distribution use. Institutional ethics policies are strictly followed.

6.4 Continuous Improvement

To remain adaptive, the system will incorporate:

- Rulebase updates for new content formats (Liu et al., 2021);
- Human-in-the-loop validation;
- Periodic reviews to ensure ethical and regulatory alignment (Zhao et al., 2019).

This ensures innovation proceeds responsibly, maintaining transparency and user protection.

Reference

Wang, Y. (2024). Survey for detecting AI-generated content. Advances in Engineering Technology Research, 11(1). https://doi.org/10.56028/aetr.11.1.643.2024

Singh, B., & Sharma, D. K. (2022). Predicting image credibility in fake news over social media using multi-modal approach. Neural Computing and Applications, 34, 21503–21517. https://doi.org/10.1007/s00521-021-06086-4

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2185–2194. https://openaccess.thecvf.com/content/CVPR2021/html/Zhao_Multi-Attentional_Deepfake_Detection_CVPR_2021_paper.html

Stephen, A. T. (2016). The role of digital and social media marketing in consumer behavior. Current Opinion in Psychology, 10, 17–21. https://doi.org/10.1016/j.copsyc.2015.10.016

Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Chua, M. C. H., & Liu, M. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. Information Processing & Management, 56(6), 102097. https://doi.org/10.1016/j.ipm.2019.102097

Liao, Q. V., Wallace, M. L., & Hines, K. (2024). AI content detection in the emerging information ecosystem. Ethics and Information Technology. https://doi.org/10.1007/s10676-024-09795-1

Jahan, S., Awan, M. J., Aslam, M. N., & Alharbi, A. (2022). Detection and moderation of detrimental content on social media platforms. Social Network Analysis and Mining, 12(1), 1–17. https://doi.org/10.1007/s13278-022-00951-3

Khan, R. A., Naseem, U., Nawaz, R., & Razzak, I. (2021). Predicting image credibility in fake news over social media using multi-modal approach. Neural Computing and Applications, 34(3), 2241–2254. https://doi.org/10.1007/s00521-021-06086-4

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), 4171–4186. https://doi.org/10.18653/v1/N19-1423

Liu, Y., Lin, H., Wu, Y., Wang, W., Xu, J., & Zhang, M. (2021). Fake news detection via multi-modal topic memory network. In Proceedings of the 29th ACM International Conference on Multimedia (ACM MM), 4367–4375. https://doi.org/10.1145/3474085.3475636

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (ICML), 6105–6114. https://proceedings.mlr.press/v97/tan19a.html

Wang, Y. (2024). Survey for detecting AI-generated content. Advances in Engineering Technology Research, 11(1), 643. https://doi.org/10.56028/aetr.11.1.643.2024

Singh, B., & Sharma, D. K. (2022). Predicting image credibility in fake news over social media using multi-modal approach. Neural Computing and Applications, 34, 21503–21517. https://doi.org/10.1007/s00521-021-06086-4

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2185–2194). https://openaccess.thecvf.com/content/CVPR2021/html/Zhao_Multi-Attentional_Deepfake_Detection_CVPR_2021_paper.html

Stephen, A. T. (2016). The role of digital and social media marketing in consumer behavior. Current Opinion in Psychology, 10, 17–21. https://doi.org/10.1016/j.copsyc.2015.10.016

Zhao, Z., Zhu, H., Xue, Z., Liu, Z., Tian, J., Chua, M. C. H., & Liu, M. (2019). An image-text consistency driven multimodal sentiment analysis approach for social media. Information Processing & Management, 56(6), 102097. https://doi.org/10.1016/j.ipm.2019.102097

Liao, Q. V., Wallace, M. L., & Hines, K. (2024). AI content detection in the emerging information ecosystem. Ethics and Information Technology. https://doi.org/10.1007/s10676-024-09795-1

Jahan, S., Awan, M. J., Aslam, M. N., & Alharbi, A. (2022). Detection and moderation of detrimental content on social media platforms. Social Network Analysis and Mining, 12(1), 1–17. https://doi.org/10.1007/s13278-022-00951-3

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

Schröer, C., Hufnagl, D., & Koch, J. (2021). A systematic literature review on applying CRISP-DM process model. In Proceedings of the 54th Hawaii International Conference on System Sciences (pp. 1666–1675). https://doi.org/10.24251/HICSS.2021.201

Van der Aalst, W. M. P., Adriansyah, A., & van Dongen, B. F. (2015). Processes meet big data: Connecting data science with process science. In Proceedings of the 9th International Conference on Business Process Management (pp. 2–7). Springer. https://doi.org/10.1007/978-3-642-32885-5_1

Floridi, L., Taddeo, M., & Turilli, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2083), 20160360. https://doi.org/10.1098/rsta.2016.0360

Majeed, A., Rana, O. F., & Rezgui, Y. (2017). Anonymization of PII using diversity-aware frameworks. Journal of Network and Computer Applications, 86, 1–12. https://doi.org/10.1016/j.jnca.2017.03.004