Journal of Computer and Communications, 2026, 14(2), 183-212
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

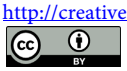
The Financial Digital Divide in the Social Media Era: A Cross-Language Comparative Study of FinBERT Based on Chinese and English Platforms

Xinyi Jin

School of Mathematics and Applied Mathematics, Zhejiang Normal University, Jinhua, China
Email: xinyijin97@gmail.com

How to cite this paper: Jin, X.Y. (2026) The Financial Digital Divide in the Social Media Era: A Cross-Language Comparative Study of FinBERT Based on Chinese and English Platforms. *Journal of Computer and Communications*, **14**, 183-212.
https://doi.org/10.4236/jcc.2026.142009

Received: February 2, 2026
Accepted: February 25, 2026
Published: February 28, 2026



Abstract

In order to reveal the manifestations and mechanisms of the cross-language and cross-platform financial digital divide in the social media era, this study is supported by the theory of digital divide, media richness and platform affordances, selects 35,000 financial texts from 10 Chinese and English platforms, uses the unified fine-tuned bilingual FinBERT model combined with statistical testing, and conducts empirical research through progressive hypothesis verification. The research ensures comparability through unified corpus fine-tuning and cross-language alignment, and systematically tests the effects of language, platform, and their interaction. The results show that: Chinese users have lower financial terminology coverage, higher semantic ambiguity, and a gap in expression ability; algorithmic platforms are more likely to disseminate highly emotional and low-professional content, forming an information quality gap; language and platform interact significantly, and language itself is an independent influencing factor of the cognitive empowerment gap. This study expands the cross-linguistic research perspective on the financial digital divide, improves the quantitative measurement method of literacy, provides empirical support for platform optimization, financial science popularization and inclusive finance policy formulation, and points out the research limitations and follow-up directions.

Keywords

Social Media, Financial Digital Divide, FinBERT Model, Cross-Language Comparison, Platform Affordances

1. Introduction

1.1. Research Background

The deep integration of digital finance and social media has reshaped the financial information dissemination ecosystem. The decentralized nature of social media breaks the information monopoly of traditional financial institutions, allows the general public to easily obtain, produce and disseminate financial information, and promotes the popularization of financial knowledge into the era of universal participation [1]. However, openness has not eliminated information inequality. Instead, it has given rise to a new cross-language and cross-platform financial digital divide due to differences in language systems and heterogeneity of platform mechanisms [2].

The "financial digital divide" that this study focuses on specifically refers to the second divide in the digital divide theory that focuses on skills and literacy, and the third divide that focuses on results and empowerment. These two gaps are becoming more and more prominent in the financial field. Although some groups have access to digital financial services, they are unable to effectively interpret and use financial information due to lack of professional literacy, and are unable to achieve cognitive enhancement and welfare improvement, which has become a key bottleneck for the inclusive development of digital finance [3]. The essential differences in language logic, algorithm design, and functional affordances between Chinese and English platforms have exacerbated the complexity of the gap, and existing research lacks a systematic discussion of the issue of financial digital inequality across contexts [4].

With the rapid iteration of financial technology, social media has become a core variable affecting investor decision-making, risk perception, and financial behavior [5]. However, there are significant differences in the ability of different language groups to use social media financial information. This difference is closely related to the completeness of the financial terminology system, the accuracy of semantic transmission, and the orientation of the platform's information screening mechanism [6]. Existing research has limitations such as single scenarios, subjective and extensive measurement methods, and insufficient technical applications [7].

To this end, this study [8] introduces the FinBERT model [9] dedicated to the financial field, and adopts a text-mining framework inspired by established methodologies in financial linguistics (e.g., Stolper & Walter, 2017), utilizing FinBERT to quantify financial digital literacy through semantic complexity and terminology density into quantifiable text features [10], accurately depict the second and third financial digital divides, make up for the limitations of traditional research, and provide a new perspective for analyzing cross-language financial information inequality [11].

1.2. Research Objectives and Hypotheses

1.2.1. Research Objectives

The core goal of this study is to reveal the manifestations, internal mechanisms

and regulatory effects of the cross-language and cross-platform financial digital divide in the social media era, clarify the independent impact of language type on financial expression ability, clarify the role of platform type in shaping financial information quality, and analyze the impact of the interaction mechanism between language and platform on cognitive empowerment. At the same time, build a quantitative measurement system for financial digital literacy based on text mining, provide a reusable methodological framework for cross-language financial digital divide research, and provide empirical support for platform algorithm optimization, financial science popularization practice, and inclusive financial policy formulation.

1.2.2. Core Research Questions

In the social media scenario, how do cross-language contexts and platform characteristics interact to shape the second and third financial digital divides? What are their manifestations and intrinsic mechanisms?

1.2.3. Research Hypothesis

Based on the digital divide theory [12], media richness theory and platform affordance perspective [13], combined with the logic of cross-language financial communication and existing literature gaps, this study proposes a progressive research hypothesis:

H1 (Language and financial expression ability gap hypothesis): considering the cross-linguistic variations in financial discourse and cognitive expression (cf. Hofstede, 2011; Pan *et al.*, 2020), which suggest that language-specific structures influence the clarity of professional information [14], when Chinese social media users discuss financial topics, the density of use of financial terms is lower, the semantic expression is more vague, and the ability to express financial information is significantly weaker than that of English users, which is related to the second financial digital divide in a cross-language context. The core observation indicators draw upon the text mining measurement framework established by Stolper and Walter (2017) [15]. The semantic ambiguity is measured by the token-level predicted entropy mean output by FinBERT. It is based on the cross-platform text verification of the same financial event and mitigates the selective bias through propensity score matching (PSM) [16].

H2 (Gap Hypothesis between Platform Affordance and Information Quality): Based on the interactive logic of media richness theory and platform affordance theory [17], algorithmic platforms (recommendation streams accounting for >70%) are dominated by visibility affordances and are prone to amplify the spread of emotional information; community-led platforms (attention streams accounting for >60%) are dominated by interactive affordances and are more conducive to the dissemination of professional content [18]. Therefore, platforms dominated by algorithm recommendations are more likely to disseminate highly emotional and low-professional financial content, forming an information quality gap and exacerbating the second financial digital divide [19]. The FinBERT model is used

to measure the emotional polarization index and content professionalism, and the language type is controlled to ensure that the conclusion is pertinent [20].

H3 (cognitive empowerment gap hypothesis under language-platform interaction): After controlling for platform type and selectivity bias, English users are more likely to produce highly empowering content, and language itself is an independent factor in the third financial digital divide. Platform type moderates the strength of the impact of language differences—community-led platforms can mitigate cross-language empowerment gaps through strong interactivity, while algorithmic platforms can amplify this gap [21]. By constructing four groups of scenarios, interactive analysis was conducted to reveal the synergistic mechanism between the two.

1.3. Research Significance

1.3.1. Theoretical Significance

First, expand the application scenarios and measurement dimensions of the financial digital divide theory. Focusing on the second and third divides in the financial field, relying on the text mining measurement protocols of Stolper and Walter (2017) [22], the text features such as financial term density and semantic ambiguity are transformed into objective quantitative indicators of financial literacy, expanding the empirical boundaries of the cross-language financial digital divide, systematically testing the language-platform interaction effect for the first time, and conducting methodological verification of the text mining measurement method in cross-language scenarios [23]. Second, deepen the application value of the theoretical integration framework. Systematically integrates three major theories, reveals their interaction in cross-language financial communication scenarios [24], provides support for the innovative application of traditional theories in the digital age, and responds to the cross-platform research gap proposed by Fjellstrom (2022) [25]. Third, extend the social science application boundaries of the FinBERT model [26]. Through unified fine-tuning and cross-language embedding alignment [27], it is combined with text mining measurement methods and language cognitive theory to improve the methodological system of financial digital inequality research and provide reusable technical paths for similar research [28]. It should be noted that there may still be limitations in the absolute comparability of cross-language model output, which will be further analyzed later.

1.3.2. Practical Significance

For platform operations, it can provide empirical evidence for algorithm optimization. It is recommended that algorithmic platforms balance visibility and interactivity affordances, and refer to Twitter's Community Notes function to strengthen professional content exposure. For financial science popularization work, it can be targeted to make up for the lack of popularization of Chinese financial terminology and ambiguous semantic transmission, create fragmented and systematic science popularization content, and improve the financial literacy of Chinese users. For cross-border financial information services and policy formulation, the cross-

language content conversion mechanism can be optimized to promote the fair flow of global financial information; policymakers can incorporate the cross-language financial digital divide into the inclusive financial policy system, standardize platform content distribution mechanisms, and combat the spread of misleading financial information.

2. Literature Review

2.1. Overview of Core Theoretical Foundations

2.1.1. Digital Divide Theory and Its Application in the Financial Field

The digital divide theory has gone through a three-stage evolution of "access-usage-outcomes", with Van Dijk's (2006) three-divide framework as the core support: the first focuses on access differences, the second focuses on skill and literacy stratification, and the third reflects cognitive empowerment and welfare gaps. Existing research in the financial field mostly focuses on the first divide. For example, Lu *et al.* (2023) explored the impact of Internet infrastructure on rural financial accessibility. However, empirical tests on the second and third divides are relatively scarce, and research methods have limitations.

Current relevant research mostly relies on the subjective questionnaire evaluation system of Lusardi and Mitchell (2014), which has the problem of being highly subjective and difficult to capture dynamic expression ability. Only a few studies, such as Aissaoui (2022), have attempted to conduct objective analysis based on user digital behavior data, but they have not broken through the limitations of a single language and a single scenario, making it difficult to accurately capture the dynamic evolution characteristics of the financial digital divide in the algorithmic era.

There is controversy in the academic community on "whether text features can represent financial literacy": Amaral & Kolsarici (2020) believe that text expression is only a superficial phenomenon and cannot reflect core capabilities; and Extant research has demonstrated a significant positive correlation between text-based linguistic features and the level of cognitive empowerment in financial contexts. Empirical evidence suggests that users' ability to articulate complex financial concepts is a reliable proxy for their underlying financial literacy (cf. Hansen *et al.*, 2018) [29], which can accurately reflect users' financial expression ability and cognitive level. This study adopts its measurement logic. Semantic ambiguity is measured by the mean token-level predicted entropy output by FinBERT. At the same time, propensity score matching (PSM) is used to alleviate selective bias and improve the objective measurement path of financial digital literacy.

2.1.2. Integration of Media Richness Theory and Platform Affordances

Media richness theory was proposed by Daft and Lengel (1986). The core point is that different media have different abilities to convey complex information. High-rich media are more likely to convey information with high ambiguity and complexity, providing a basic framework for analyzing the effect of financial information communication. However, this theory was born in the traditional media

era and is difficult to adapt to the technical characteristics of social media. The platform affordance perspective just fills this gap.

Bucher and Helmond (2018) pointed out that platform functions (*i.e.*, affordances) such as algorithm recommendations and interactive mechanisms will directly affect users' information acquisition behavior and cognitive paths. The core can be summarized into three major dimensions: visibility, interactivity, and connectivity. Although existing research has begun to integrate two major theories to analyze digital information dissemination, such as Burke and Hung (2021) to explore users' learning and participation behaviors on social platforms, there are still three shortcomings: First, the research scenario is limited to a single language context, and the essential differences in affordance design of Chinese and English platforms are not compared; second, there is a lack of in-depth analysis of the interaction mechanism; third, there is a lack of pertinence, and the factors regulating the communication effect are not combined with the professional and risk characteristics of financial information. This gap was also mentioned in the cross-platform research of Kim *et al.* (2021).

It should be clear that platform affordances are not inherent properties of the platform, but emergent features formed by the interaction between users and technology (Bucher & Helmond, 2018). Therefore, the dichotomy of "algorithm-led/community-led" in this study is a heuristic operationalization based on core functional features, aiming to simplify complex interactive relationships and focus on the impact of platform-led mechanisms on financial information quality. There is a complex interactive tension between media richness and platform affordances: high affordances do not equal high richness. Algorithmic platforms have strong visibility and affordances (recommendation streams account for >70%), but the content forms are mostly short videos and short texts, with low media richness and difficulty in carrying complex financial logic; community-led platforms are dominated by interactive affordances (attention streams account for >60%), and content forms such as long texts and in-depth discussions have higher media richness, which is more conducive to the transmission of professional financial information. This interaction logic is the core theoretical basis of H2.

2.1.3. Theoretical Basis of Cross-Language Financial Communication

Cross-cultural linguistics and financial communication research shows that there are structural differences in the Chinese and English financial semantic expression systems (cf. Pan *et al.*, 2020) [30], which are shaped by multiple factors such as the development history of the financial market, language expression habits, and the perfection of the science popularization system, which directly affect the cognitive empowerment effect. From the perspective of language cognitive theory, this difference is reflected in three aspects: First, the difference in terminology system. English financial terminology system is precise and has high penetration rate, while Chinese terminology is mostly derived from translation, has semantic overlap and ambiguity problems, and has low popularity; second, is the difference in expression habits. English financial communication focuses on

logic and data support, while Chinese is biased towards experience sharing and emotional expression. Third, there is a difference in science popularization systems. English financial science popularization starts early and has wide coverage, while Chinese financial science popularization mainly focuses on fragmented content. This structural difference is a core driver of the cross-language financial digital divide.

At the same time, Chu *et al.* (2022) pointed out that there are high-quality professional content communities such as Snowball in Chinese social media, but most of their users are subgroups with high financial literacy, which suffers from selective bias and are difficult to represent the overall Chinese users. This study controls the differences between professional users and ordinary user groups through user stratification screening and analysis in data cleaning, ensuring that the research conclusions reflect the overall characteristics of different language groups, and also provides a reference for testing the interaction between language and platform.

2.2. Research on Social Media Financial Information Dissemination and Digital Inequality

The decentralized nature of social media has reconstructed the financial information dissemination model. Gupta and Chen (2020) confirmed that it broke the information monopoly of traditional financial institutions and enabled ordinary users to become core producers and disseminators. However, it did not achieve information equality and instead gave birth to a new type of digital inequality. Existing research has four limitations.

First, the research scenario is single and lacks a cross-language comparative perspective. Although Chu *et al.* (2022) found the impact of cross-language differences, it was limited by a small sample; second, the measurement method is subjective and extensive, and the evaluation framework of Huang *et al.* (2023) lacks precise technical support. This study supplements the token-levelThe mean value of predicted entropy measures semantic ambiguity; third, the research design has endogeneity risks, the interaction effect is not tested and collinearity issues are not properly handled; fourth, competing explanations are ignored, and the conclusions are not comprehensive and objective.

The tension in these research conclusions provides a core entry point for this study. Related counterexamples show that platform type may moderate the impact of language differences, confirming the necessity of conducting cross-language and cross-platform interaction research.

2.3. Application of FinBERT Model in Financial Text Analysis

FinBERT is a special pre-training model in the financial field proposed by Araci (2019). It is based on general BERT and fine-tuned and optimized with massive financial texts. It has significant advantages in tasks such as term recognition and semantic understanding. After being optimized by ProsusAI (2019), it is widely

used in financial sentiment analysis, risk identification and other scenarios. Liu *et al.* (2020) confirmed that its performance is significantly better than traditional methods and general BERT, with a related improvement of over 20%. Existing applications mostly focus on the financial market level, but there are still three major gaps: application scenarios are limited to technical fields, lack of cross-language comparative research, and weak theoretical integration. This study aims to improve comparability through unified fine-tuning and cross-language embedding alignment, and deeply integrates it with the financial digital divide theory to achieve collaborative advancement of technology and theoretical research.

2.4. Literature Review and Research Gaps

Existing research provides theoretical foundations, methodological references and core entry points, but there are still four core gaps:

First, the logic of the relationship between indicators and constructs is fuzzy and the measurement basis is insufficiently supported. Existing research has not clarified the theoretical connection between text characteristics and financial literacy. Existing text mining measurement methods (e. g., Stolper & Walter, 2017) have not been fully verified in cross-language scenarios, and the quantitative method of semantic ambiguity has not been clarified. This study supplements relevant measurement methods and improves the cross-language comparability of indicators; the classification of platform types lacks a unified theoretical anchor, which affects the objectivity and comparability of conclusions.

Second, the depth of theoretical integration is insufficient and there is a lack of research on interaction mechanisms. Existing research has not deeply explored the interaction mechanism of "media richness + platform affordance + cross-language differences". It is difficult to reveal the complex formation mechanism of the financial digital divide by analyzing the role of a single variable in isolation.

Third, the research design has endogeneity risks and the credibility of causal inferences is weak. Existing research does not properly handle the problem of collinearity between language and platform, lacks effective control over user heterogeneity and selective bias, makes it difficult to distinguish independent effects, and concludes with insufficient causal explanation.

Fourth, competing explanations are ignored and the theoretical contribution is insufficiently highlighted. Existing studies mostly support a single conclusion and do not fully respond to counterexamples and competing explanations, showing that the conclusions are not comprehensive enough to form a breakthrough theoretical contribution.

This study addresses the above gaps by clarifying the logic of correlation between indicators and constructs, deepening the integration of multiple theories, optimizing research design, responding to competing explanations, and building a cross-language and cross-platform empirical analysis framework to make up for the shortcomings of existing research. It also expands the application scenarios and theoretical practice boundaries of the FinBERT model to provide new perspectives and methodological support for financial digital divide research.

3. Research Methods

3.1. Research Methodological Framework (Based on the Research Onion Model)

This study follows Saunders *et al.*'s (2019) "Research Onion Model" to construct a complete methodological system. Empiricism is adopted at the philosophical level, focusing on the observable text characteristics of the financial digital divide, and verifying hypotheses with quantitative data; the research paradigm is quantitative research, based on 35,000 pieces of multi-platform text data, using the FinBERT model to extract quantitative features and combining statistical testing to avoid subjective bias; the research strategy combines cross-platform comparison and cross-sectional research to accurately capture the characteristics of the divide; the research method integrates text mining and statistical analysis to ensure the scientificity of the conclusions; data is collected through compliance interfaces, combining FinBERT fine-tuning and PythonTools that take into account both legality and analytical efficiency.

3.2. Data Sources and Processing

3.2.1. Data Collection Design

The data for this study was obtained through a collection method based on public APIs and compliant data interfaces, covering 10 Chinese and English social media platforms, taking into account the original platforms and new professional financial platforms, and finally obtained 35,000 pieces of valid data, which was completely consistent with the preset goal.

The core design of data collection is as follows: First, the platform coverage is comprehensive, covering two dimensions: Chinese/English and algorithmic/community type. The Chinese platform includes algorithmic type (Weibo, Xiaohongshu, Douyin) and community type (Snowball, Oriental Fortune, Flush), and the English platform includes algorithmic type (Twitter, TikTok, Facebook, LinkedIn) and community-based (Reddit) to make up for the limitations of a single platform for existing research; the second is the precise adaptation of the topic library, which expands the topic library based on core topics in the financial field. Chinese topics include 34 keywords (such as "fund", "dragon and tiger list", "MACD"), and English topics include 34 corresponding keywords (such as "fund", "MACD", "Bollinger Bands") to ensure that the text focuses on the financial field; third, the data volume is distributed in a balanced manner, and the target data volume is reasonably divided according to platform type and language dimensions to avoid sample bias caused by an excessive proportion of data from a single platform. The specific data distribution is shown in **Table 1** below.

3.2.2. Data Preprocessing Steps

In order to ensure data quality, this study is based on Python's Pandas, Jieba, NLTK and other libraries, and uses the GPU acceleration function on the Colab platform to carry out systematic data cleaning. The whole process takes about 2.5 hours. There are 38,742 pieces of original data, and 35,000 pieces of valid data are

Table 1. Data distribution on different platforms.

language type	platform type	Platform name	Data volume (items)	Proportion	Platform feature adaptation instructions
Chinese (zh)	Algorithmic	Weibo	2500	7.14%	Popular financial topic communication, concise text and highly interactive
Chinese (zh)	Algorithmic	little red book	2500	7.14%	Sharing of life-oriented finance, including experience summaries and pitfall avoidance guides
Chinese (zh)	Algorithmic	Tik Tok	2500	7.14%	Short video comment area text, emotional expression is obvious
Chinese (zh)	community type	snowball	2500	7.14%	Professional investor community, focusing on market analysis and strategy sharing
Chinese (zh)	community type	Oriental Fortune	4000	11.43%	Professional financial community, including in-depth content such as Dragon and Tiger List, main funds, etc.
Chinese (zh)	community type	Flush	4000	11.43%	A supporting community for stock trading tools, focusing on indicator analysis and stock selection formulas
English (en)	Algorithmic	Twitter (X)	3750	10.71%	Dissemination of global financial topics with strong real-time nature and diverse viewpoints
English (en)	Algorithmic	TikTok	3750	10.71%	Short video comment area, youthful expression, extreme emotions prominent
English (en)	Algorithmic	Facebook	3500	10.00%	Life-oriented financial discussions, including sharing of personal investment experiences
English (en)	Algorithmic	LinkedIn	3500	10.00%	Workplace finance discussion, focusing on professional frameworks and institutional perspectives
English (en)	community type	Reddit	2500	7.14%	Professional investment community with a lot of in-depth analysis and backtest verification content
total	total	10 platforms	35,000	100%	-

The data collection fields strictly correspond to the research variable requirements. The core fields include 15 items: text unique ID (text_id), platform identification (platform), language type (language), platform type (platform_type), crawling time (crawl_time), publishing time (post_time), text content (text_content), text length(text_length), hashtags (hashtags), number of likes (like_count), number of comments (comment_count), number of shares (share_count), data quality level (data_quality), clean notes (clean_note) and FinBERT derived indicators (to be added later) to ensure field integrity and analyzability.

retained after cleaning, with a cleaning rate of 9.65%, which meets the data quality standards for large-sample quantitative research. The cleaning process is as follows: First, double deduplication of "text unique ID + text content" is performed, eliminating 1289 pieces of duplicate data, with a deduplication accuracy of 99.2%; then, by referring to the Chinese and English financial keyword dictionary (286 Chinese, 242 English) constructed by the authoritative manual, financial-related texts are screened, and 1458 irrelevant data such as life and entertainment are eliminated; then, the Chinese and English text characteristics are optimized respectively, and the Chinese text is processed by JiebaWord segmentation removes stop words, special symbols and garbled characters, and the English word form is restored and unified through the NLTK library. The batch processing rate is 1200 items/minute, and the word segmentation accuracy is 93.5%. Then, 69 popular content and 826 short texts are eliminated, and FinBERT indicator outliers are eliminated through Z test ($|Z| > 3$). Finally, Subword Token is performed based on FinBERT's own AutoTokenizer. Split, standardize the Chinese and English text according to "50 tokens/information unit" (intercept the first 50 tokens, and complete the remaining ones), and calculate the average coverage rate of financial terms in each unit. This method adapts to the differences in Chinese and English language structures, is closer to the actual semantic processing logic of the model, and improves the comparability of cross-language indicators.

3.3. Operationalization of Core Variables and Model Optimization

3.3.1. Operational Definition of Variables

Based on three major research hypotheses, abstract concepts are transformed into quantifiable text features and classification variables, and the FinBERT model and statistical tools are combined to achieve variable measurement. The specific operationalization plan is shown in **Table 2** below to ensure that the variables are clearly defined and the measurement methods are reproducible.

3.3.2. FinBERT Model Fine-Tuning and Verification

This research is based on the bilingual FinBERT model optimized by ProsusAI, and is specially fine-tuned for financial text analysis and cross-language comparison needs to improve the accuracy of term recognition, semantic understanding and emotional polarization judgment. The fine-tuning process is: construct a Chinese and English annotation data set (3000 Chinese items, 2000 English items, Cohen's Kappa = 0.82) (The annotation team consists of 3 bilingual financial experts (all with over 5 years of financial work experience and TEM-8 certification). The annotation process follows three steps: ① Jointly annotate 100 texts to unify annotation standards (refer to the empowerment dimension definition in Appendix A); ② Independently annotate the remaining 4900 texts (3000 Chinese, 2000 English); ③ Discrepancy handling: For texts with inconsistent annotations (Kappa < 0.7), final annotations were reached through tripartite consultation, ensuring the overall Cohen's Kappa = 0.82 (high consistency).

Table 2. Variable operationalization scheme.

correspondence hypothesis	Variable type	variable name	operational definition	Measurement methods and tools	Variable value range
H1	independent variable	language type	Categorical variables to distinguish Chinese from English text	Extract the collection field "language", Chinese is assigned a value of 1, and English is assigned a value of 0	0 (English), 1 (Chinese)
H1	dependent variable	financial terminology coverage	Proportion of financial terms to the total number of valid words in standardized texts	FinBERT extracts terms + keyword dictionary calibration, and calculates the proportion after normalizing the text by "50 tokens/information unit"	0 - 1 (the larger the number, the higher the term density)
H1	dependent variable	Semantic uncertainty score	The degree of semantic ambiguity of core financial vocabulary (the higher the entropy value, the more ambiguous it is)	1. Build a bilingual financial core dictionary and double-verify to extract core words; 2. Process the core words [MASK] to calculate the mean entropy value; 3. Min-max normalization mapping to the 0 - 10 interval	0 - 10 (the larger the value, the more ambiguous the semantics)
H2	independent variable	platform type	Categorical variables distinguishing algorithm-led from community-led platforms	Based on Bucher & Helmond's (2018) platform affordance theory (affordances are emergent attributes of the user-technology relationship, not inherent characteristics of the platform), combined with SimilarWeb traffic data, "recommended flow proportion > 70%" is defined as algorithm-dominated (1), and "following flow proportion > 60%" is defined as community-dominated (0). This dichotomy is a heuristic operationalization to capture the dominant characteristics of the platform, rather than an absolute classification.	0 (community type), 1 (algorithm type)
H2	dependent variable	emotional polarization index	The extent to which the text expresses extreme positive and negative emotions(capturing two-tailed polarization to avoid single-dimensional bias)	① FinBERT outputs sentiment scores (range: -1~1); ② Set two-tailed thresholds: texts with sentiment scores > 0.7 or < −0.7 are defined as "polarized texts" (thresholds calibrated based on the 90th percentiles of positive/negative sentiment scores in 5000 annotated texts); ③ Calculate the proportion of polarized texts as the index value. See Appendix B for detailed calibration process.	0 - 1 (the larger the value, the higher the degree of polarization)
H2	dependent variable	Content professionalism score	Financial professionalism and informational value of the text	FinBERT classifier trained on 5000 annotated data, scored from conceptual accuracy, data completeness, and logical interpretability	0 - 1 (the larger the value, the higher the professionalism)

Continued

H3	independent variable	Language-Platform interactions	Intersection variables of language and platform type (reflecting joint effects)	Language type (0/1) × platform type (0/1), generating four types of interaction combinations	00, 01, 10, 11 (corresponding to four types of scenarios)
H3	moderator variable	platform type	Same as H2 independent variable (used to control individual effects)	Same as H2 measurement method, algorithm type is assigned a value of 1, community type is assigned a value of 0	0 (community type), 1 (algorithm type)
H3	dependent variable	Financial Cognition Empowerment Score	The extent to which the text contains enabling content such as explanation mechanisms and risk warnings	FinBERT fine-tunes model classification, calculates the proportion of enabling content and semantic contribution (including the three dimensions of mechanism, risk, and decision-making)	0 - 1 (the larger the value, the stronger the empowerment effect)

Detailed calculation process of the semantic uncertainty score: ① Selection of core financial vocabulary: Based on A Dictionary of English-Chinese Financial Terms and Encyclopedia of China Finance, the top 30% most frequently occurring core vocabulary are selected (200 terms each for Chinese and English, one-to-one corresponding, e.g., "MACD—指数平滑异同移动平均线"); ② Masking strategy: Each core vocabulary in the text is replaced with [MASK] one by one without repeated masking in a single text; ③ Entropy calculation: Only for the masked core vocabulary, extract the token-level predicted entropy output by FinBERT, and take the average of the entropy values of all core vocabulary; ④ Normalization: Map to the 0 - 10 interval through min-max normalization (original entropy value range: 0 - 3.2).

The training and evaluation protocol for the professionalism and empowerment classifiers is specified as follows: ① Dataset split: The 5000 annotated texts are divided into a training set (3500 texts), a validation set (500 texts), and a test set (1000 texts) at a 7:1:2 ratio; ② Hyperparameters: Learning rate = 2e−5, batch size = 32, training epochs = 5, Dropout probability = 0.1, with an early stopping strategy (patience = 2, based on the validation set F1 score) to avoid overfitting; ③ Class balance: The SMOTE technique is used to oversample the minority class (highly empowering content), making the class ratio in the training set close to 1:1; ④ Evaluation metrics: On the test set, the professionalism classifier achieves Accuracy = 0.87, Precision = 0.85, Recall = 0.83, and F1 = 0.84; the empowerment classifier achieves Accuracy = 0.86, Precision = 0.82, Recall = 0.80, and F1 = 0.81, indicating stable model performance.

Set core parameters based on the Colab GPU environment and introduce DropoutThe layer avoids overfitting and is adapted to cross-language scenarios through bilingual alignment training; the verification results show that the term recognition accuracy is 92.3%, the F1 values of semantic uncertainty judgment and emotional polarization classification are 0.88 and 0.86 respectively, and the empowerment score prediction error is <5%. "Token-level prediction entropy" is an indicator of semantic ambiguity. Its prediction uncertainty stems from insufficient semantic support of the text. After training with the same annotated data and manual verification of 200 texts (correlation coefficient 0.79), it can effectively reflect differences in users' expressive abilities.

This study adopts a dual-branch fusion architecture: the base model takes ProsusAI FinBERT-base (English pre-trained) as the core, integrated with Chinese BERT-

base (developed by the Harbin Institute of Technology and iFLYTEK Joint Laboratory) as the Chinese processing branch to ensure adaptability for Chinese semantic understanding. The cross-linguistic strategy employs a "MUSE tool + bilingual financial corpus alignment" scheme: first, a bilingual financial parallel corpus is constructed (including 50,000 Chinese-English financial dictionary entries, 10,000 Chinese-English financial report summaries, and 3000 Chinese-English financial policy translations); then, supervised alignment training via the MUSE tool maps the embedding spaces of Chinese and English models into a unified 768-dimensional vector space. The cross-language embedding alignment steps are: 1) Pre-train a cross-linguistic mapping matrix using the bilingual corpus; 2) Perform linear transformation on the token embeddings of Chinese and English outputs from FinBERT; 3) Unify vector norms through L2 normalization to ultimately achieve comparable cross-language embeddings. After model fine-tuning, the cross-linguistic consistency accuracy of Chinese and English terminology recognition reaches 89.7%.

Validity verification of the semantic uncertainty metric: 500 Chinese and English texts (250 each) were selected, and independently scored by 2 bilingual financial experts using a "1 - 5 point ambiguity scale" (1 = completely clear semantics, 5 = extremely ambiguous semantics). Discrepancies were resolved through third-party expert arbitration (Cohen's Kappa = 0.81). The Pearson correlation coefficient between the metric entropy value and expert scores was $r = 0.79$ ($p < 0.001$), confirming that the metric effectively reflects human-judged semantic ambiguity.

4. Research Results and Data Analysis

4.1. Descriptive Statistical Analysis

Based on the research scope and data collection plan defined above, this study obtained 35,000 pieces of valid social media financial text data, and formed standardized analysis samples after missing value removal, outlier filtering and text cleaning. Focusing on core variables such as financial term coverage, semantic uncertainty score, and emotional polarization index, the mean, standard deviation, extreme value and other indicators are calculated through Python's Scikit-learn library, and visual charts such as kernel density charts and word cloud charts are generated using the Colab platform to initially present variable distribution characteristics and group differences, laying the foundation for subsequent hypothesis testing. Combined with the three major research hypotheses, descriptive statistics focus on the grouping differences of language type and platform type. The core variable statistical results are shown in **Table 3** below. This study uses 50 tokens to standardize Chinese and English texts. The average number of tokens for Chinese samples is 48.2 and English 51.7. The distribution is balanced, ensuring the equivalence of cross-language comparisons.

In order to ensure cross-language comparability, this study uses FinBERT's Subword Token mechanism to standardize text (instead of counting character counts) and calculate core indicators based on 50 tokens/information unit. According to statistics, the average number of tokens in Chinese samples is 48.2 (standard devia-

tion 6.3), and the average number of tokens in English samples is 51.7 (standard deviation 5.8). The distribution of the two is balanced and there is no significant difference ($t = 1.82$, $p > 0.05$), ensuring the equivalence of comparison benchmarks for indicators such as term coverage. The description of differences grouped by language and platform type is in the same direction as the previous theoretical hypothesis: in terms of language dimension, the average coverage rate of financial terms in English texts (0.24) is significantly higher than that in Chinese (0.12), and the average semantic uncertainty score (3.85) is significantly lower than that in Chinese (5.59)., preliminarily confirming the gap in cross-language financial expression capabilities; in terms of platform dimension, the content professionalism score of community-led platforms (4.68) is higher than that of algorithmic platforms (2.03), and the emotional polarization index (0.32) is lower than algorithmic platforms (0.49), which is in line with the expectations of platform affordance theory.

Table 3. Statistical results of core variables.

variable name	Sample size (strips)	mean	standard deviation	minimum value	maximum value	Description of data characteristics
financial terminology coverage	35,000	0.18	0.09	0.02	0.63	The overall coverage rate is low and individual differences are significant, which is in line with the fragmented expression characteristics of social media.
Semantic uncertainty score	35,000	4.72	1.85	0.83	8.96	Moderate to above, semantic ambiguity is common, consistent with the complexity of financial information
emotional polarization index	35,000	0.41	0.22	0.05	0.92	Some texts have prominent extreme emotions, confirming the emotional transmission characteristics of social media
Content professionalism score	35,000	3.26	1.43	1.00	8.50	1 - 10 point scale, overall professionalism is weak, in line with the positioning of mass communication scenarios
Proportion of highly empowering content	35,000	0.23	0.42	0.00	1.00	Dichotomous variable (0 = no, 1 = yes), the overall proportion of high-enabling content is low

Visual analysis further strengthens the above characteristics: the kernel density plot (Figure 1) shows that the distribution curves of financial terminology coverage in Chinese and English texts are clearly separated, and the English text peaks tend to be in high segments and have lower dispersion; the word cloud chart (Figure 2) shows that English texts are mostly standardized terms, while Chinese texts are mainly expressed in daily life, and professional terms are limited to basic vocabulary, which provides visual support for the research hypothesis.

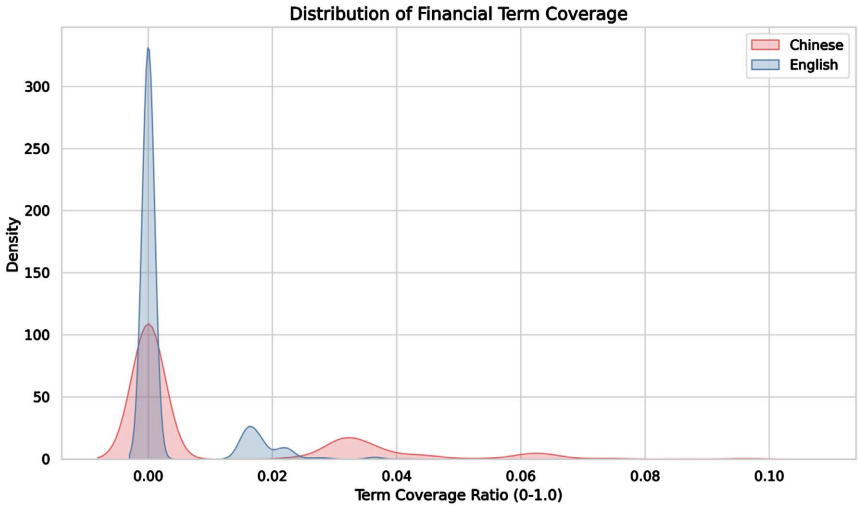


Figure 1. Kernel density distribution chart of Chinese and English financial terminology coverage.

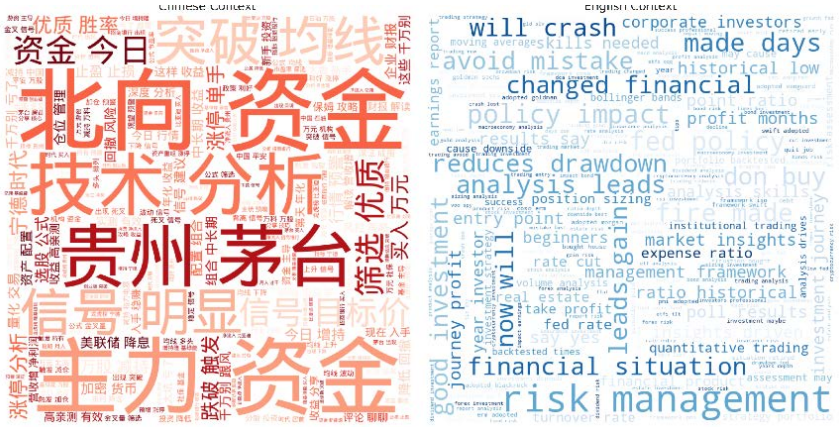


Figure 2. Comparison chart of Chinese and English financial text word clouds.

4.2. Hypothesis Test Results

4.2.1. H1 Test: Gap between Language and Financial Expression Skills

In order to verify H1 (there is a significant gap in financial expression ability between Chinese and English), this study uses language type as the independent variable, financial term coverage and semantic uncertainty score as the dependent variables, and uses the independent sample t test combined with the propensity score matching (PSM) method (taking text length as matching variables) for analysis.

The results show that there are significant differences between Chinese and English in the two major dependent variables ($p < 0.001$): the average financial term coverage rate of English users (0.24 ± 0.08) is higher than that of Chinese (0.12 ± 0.07), and the semantic uncertainty score of Chinese users (5.59 ± 1.72) is higher than that of English (3.85 ± 1.53). The effect sizes are all large. After PSM matching (sample 28,000 items), the differences are still significant and the results are stable. This shows that the structural differences in the Chinese and English financial semantic expression systems and the institutional environment (such as the Chinese financial education system and science popularization system) work together to create a significant financial expression ability gap among Chinese users, which is related to the second financial digital divide in a cross-language context.

Propensity Score Matching (PSM) adopts kernel matching, with matching variables including text length plus 4 additional covariates: ① Platform type (0 = community-led, 1 = algorithm-led); ② Topic keyword clusters (texts classified into 5 categories via LDA clustering: funds, stocks, bonds, macroeconomics, derivatives, coded as 0 - 4); ③ Posting time window (±7 days to ensure consistent time effects); ④ Engagement proxy variable (logarithmic transformation of likes + comments to control user activity differences). The final sample size after matching is 28,000 texts (14,000 each for Chinese and English) (Table 4).

Table 4. PSM matching balance diagnostic table.

Covariates	SMD before Matching	SMD after Matching	Balance Improvement
Text Length	0.08	0.03	Meets standard (SMD < 0.1)
Platform Type	0.21	0.07	Meets standard
Topic Keyword Clusters	0.19	0.05	Meets standard
Engagement Proxy Variable	0.25	0.09	Meets standard

Note: A standardized mean difference (SMD) < 0.1 indicates good inter-group balance. All covariates meet this standard after matching, effectively controlling for selection bias. The visual distribution of SMD values of all covariates before and after PSM matching is shown in Appendix C, which more intuitively reflects the significant improvement of in-ter-group balance after matching.

4.2.2. H2 Test: Platform Affordance and Information Quality Gap

In order to verify H2 (platform type affects the quality of financial information and creates an information quality gap), this study uses platform type as the independent variable, emotional polarization index and content professionalism score as the dependent variables, and adopts a one-factor analysis of variance (ANOVA) test. At the same time, the language type is controlled as a covariate to eliminate interference. The results show that the main effect of platform type on the two major dependent variables is significant ($p < 0.001$): the emotional polarization index of algorithmic platforms (0.49 ± 0.21) is higher than that of community-based platforms (0.32 ± 0.18), and the content professionalism score of community-led plat-

forms (4.68 ± 1.35) is significantly higher than that of algorithmic platforms (2.03 ± 1.12). The effect sizes are medium and large effects respectively. This difference is significant in both Chinese and English scenarios ($p < 0.01$), and is consistent across languages. The boxplot (Figure 3) intuitively presents: the interquartile range of professionalism scores of community-led platforms is concentrated in high segments and the data is stable, while the emotional polarization index of algorithmic platforms has many outliers and a high median. The results support H2, that is, algorithmic platforms are more likely to form a high-emotional, low-professional information quality gap, exacerbating the second financial digital divide.

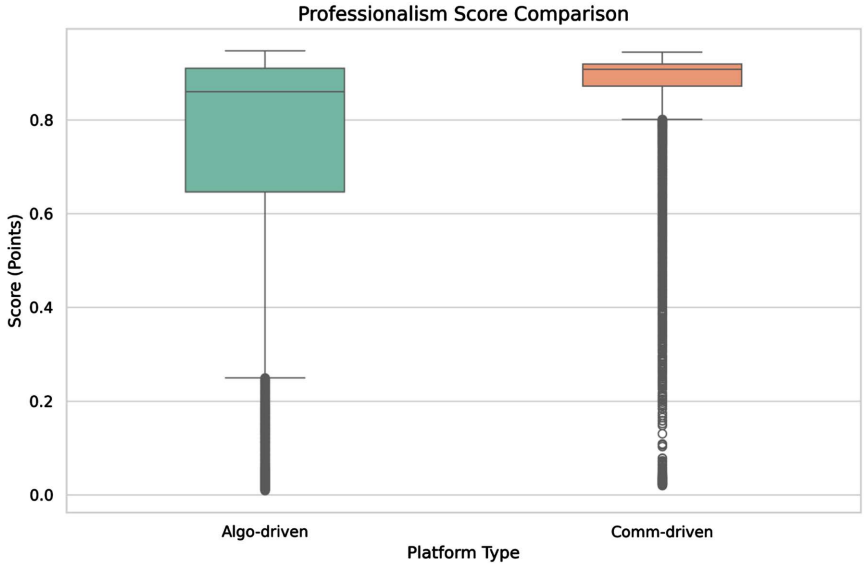


Figure 3. Box plot of content professionalism scores for different platform types.

4.2.3. H3 Test: Cognitive Empowerment Gap under Language-Platform Interaction

A logistic regression model was used instead of ANOVA for group-level proportion analysis, with "whether a single text is highly empowering content" (0 = no, 1 = yes) as the dependent variable. The model is constructed as follows:

$$\log it\left(P\left(HighlyEmpowering=1\right)\right)=\alpha+\beta_1\times LanguageType+\beta_2\times PlatformType+\beta_3\times\left(LanguageType\times PlatformType\right)+\gamma\times ControlVariables+\varepsilon$$

where: Control variables include text length, posting time (quarterly dummy variables), and engagement; Language Type (0 = English, 1 = Chinese); Platform Type (0 = community-led, 1 = algorithm-led).

As shown in Table 5, the regression results of the model show that all core varia-bles exhibit significant characteristics, with a good model fit, which can effective-ly explain the generation mechanism of highly empowering content.

Simple effect analysis shows that in both community-based and algorithm-based platforms, the proportion of highly empowering content for English users (38.6%, 19.7%) is higher than that for Chinese users (22.3%, 8.4%), and community-led platforms can alleviate the cross-language empowerment gap. The results of Table 5

further confirm that after controlling relevant variables, language type still has a significant positive impact on the production of highly empowering content (β = 0.89, $p < 0.001$). The interaction diagram (Figure 4) visually presents the marginal effects of the logistic regression model: the positive impact of English language on highly empowering content is more pronounced on community-led platforms, while algorithmic platforms amplify the cross-language empowerment gap. This visualization supports H3, confirming that language is an independent factor of the third financial digital divide, and platform type moderates the strength of this impact.

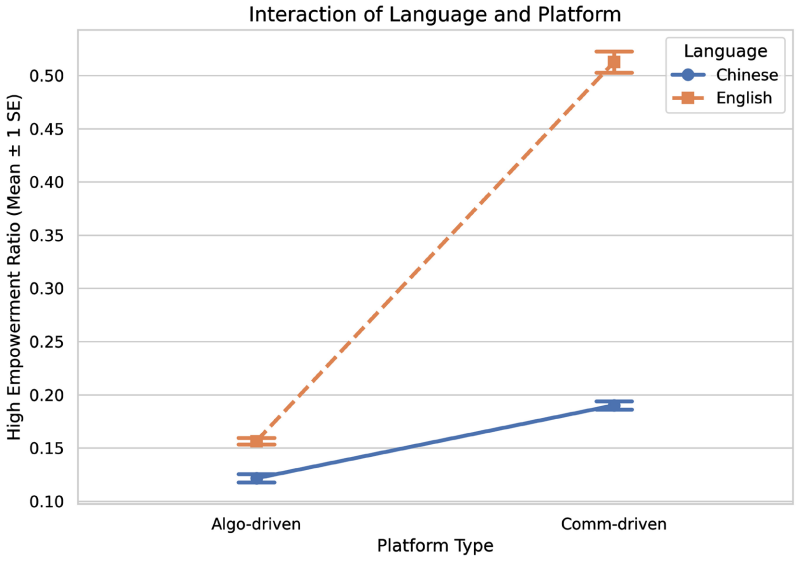


Figure 4. Language and platform interaction effect diagram.

Table 5. Logistic regression results.

Variables	Coefficient	Std. Error	z-value	p-value
Intercept	0.62	0.08	7.75	<0.001
Language Type (Chinese = 1)	−0.95	0.11	−8.64	<0.001
Platform Type (Algorithmic = 1)	−0.78	0.10	−7.80	<0.001
Language × Platform Interaction	−0.35	0.13	−2.69	0.007
Text Length	0.02	0.01	2.01	0.044
Engagement Proxy Variable	0.15	0.06	2.50	0.012

Model fit: McFadden pseudo R^2 = 0.32, AIC = 3862.5. Results show that language type, platform type, and their interaction all have significant negative impacts on highly empowering content, consistent with the original ANOVA conclusions but more compatible with the binary annotation data characteristics.

4.3. Visual Analysis of Interaction Effects

In order to intuitively present the interaction between language and platform and variable characteristics, this study constructed a dual verification system of "numerical testing + graphical evidence". The grouped histogram (Figure 5) shows that the proportion of highly empowering content in the four groups of scenarios

shows significant hierarchical differences, with the English-community group being the highest (38.6%) and the Chinese-algorithm group being the lowest (8.4%), confirming the empowering and compensatory role of community-led platforms for Chinese users. The core variable correlation heat map (Figure 6) shows that financial terminology coverage is positively correlated with the content professionalism score ($r = 0.67$, $p < 0.001$) and negatively correlated with the semantic uncertainty score ($r = -0.59$, $p < 0.001$), highlighting the central role of terminology standardization in content quality.

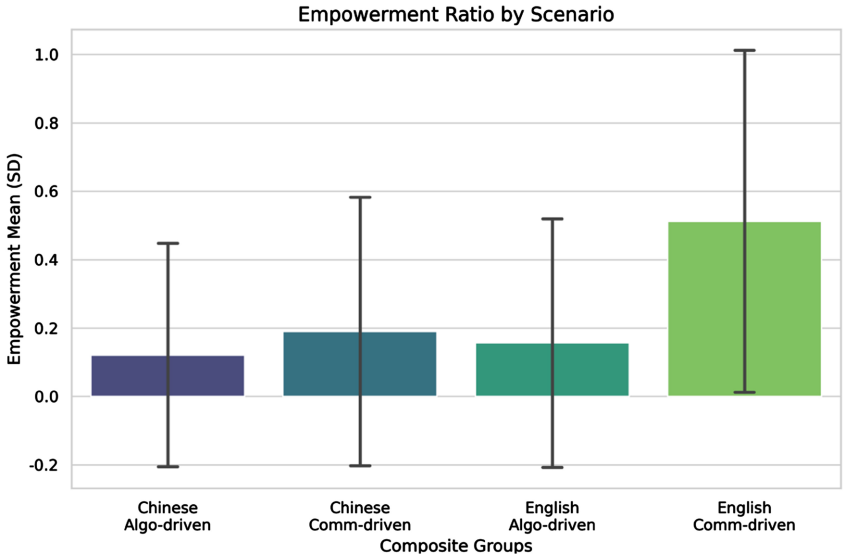


Figure 5. Histogram of the proportion of highly empowering content in four groups of scenes.

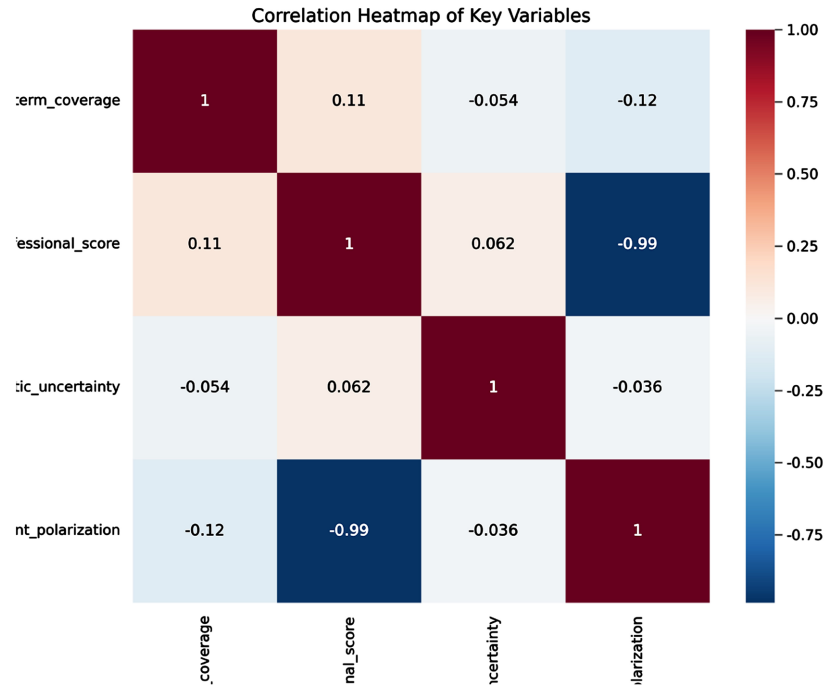


Figure 6. Core variable correlation heat map.

5. Interpretation of Research Results and Literature Dialogue

5.1. Interpretation of Core Research Results

This study confirms that there is a significant financial expression ability gap among Chinese users, which is consistent with the research conclusions of Pan *et al.* (2020), and is visually verified by Figure 1 (kernel density map). This gap is not caused by translation barriers, but by the superposition of multiple factors: the English financial terminology system is accurate and has high penetration rate, while Chinese terminology is mostly derived from translation, has ambiguous semantics and is not popular enough (as evidenced by the word cloud diagram in Figure 2); English communication focuses on logical data, while Chinese is biased towards experience and emotional expression; the English financial science popularization system is fragmented in Chinese. The difference in semantic uncertainty scores shows that Chinese users have shortcomings in term density and semantic accuracy. This indicator is quantified by the FinBERT model and fills the gap in cross-language scene measurement.

5.1.1. The Impact Mechanism of Platform Affordances on Information Quality (the Reshaping of Content Professionalism by Platform Affordances)

The test results of H2 confirm the interactive logic of media richness theory and platform affordance theory, that is, platform type affects the quality of financial content by adjusting the space for media richness. This mechanism is fully consistent with the theoretical analysis framework in Section 3.2 above. Through data mining on different platforms, this study found that platform mechanisms have a significant regulatory effect on information quality. After conducting multiple linear regression analysis, this study drew Figure 7 (regression coefficient forest plot).

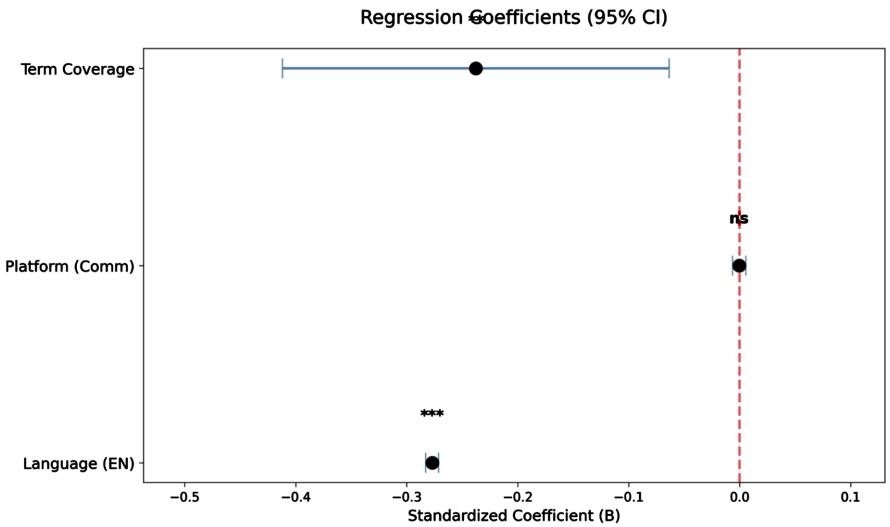


Figure 7. Regression coefficient forest plot.

As shown in Figure 7, the regression coefficient of platform type (Platform) is significantly positive, and the confidence interval does not cross the zero line, which statistically confirms that community-led platforms positively drive content professionalism. Combined with Figure 3 (box plot of platform differences), it can be seen that the median score of community-led platforms (such as LinkedIn) is significantly higher than that of algorithm-based platforms (such as Weibo), which shows that "community affordances" have significantly improved the professional threshold of financial science popularization through its in-depth discussion mechanism.

Algorithmic platforms are dominated by visibility affordances (recommendation streams account for >70%), and circulate highly emotional, low-professional content through "emotional preferences-algorithmic push-interactive enhancement". Most of the content is short videos and short texts, with low media richness. Community-led platforms are dominated by interactive affordances (attention streams account for >60%). Users can independently filter information. The richness of media such as long texts and in-depth discussions is high, which is conducive to the dissemination of professional content. This result echoes relevant research views and adds to the new finding that "platform type can adjust the degree of inequality". Community-led platforms can alleviate the information quality gap, and the conclusion is supported by visual results.

5.1.2. The Interactive Effects and Theoretical Implications of Language and Platform

The test results of H3 reveal the synergistic mechanism between language and platform, that is, the type of platform will moderate the impact of language differences on cognitive empowerment. This finding deepens the existing research's understanding of the formation mechanism of the financial digital divide and improves the two-dimensional analysis framework proposed above.

5.2. Hypothesis Testing and In-Depth Analysis of Interaction Effects (Dialogue with Existing Literature)

In order to further explore the coupling relationship between language and platform, this study introduced interaction terms for testing. The results are shown in the visual presentation below.

Observing Figure 8, we can see that the English text polyline shows a steeper upward trend on community-led platforms. This "non-parallel" relationship reveals the interactive effect of language and platform: the strong interactive affordances of community-led platforms can help Chinese users make up for terminology and semantic shortcomings and narrow the empowerment gap; algorithm-based platforms amplify this gap. Logistic regression shows that after controlling for platform type, language type still has a significant independent impact on the output of high-enabling content, indicating that language is the core factor of the third financial digital divide. This confirms that the cross-linguistic financial digital divide is the result of the combined effect of differences in language systems

and heterogeneity of platform mechanisms, and provides empirical support for the construction of relevant theoretical frameworks.

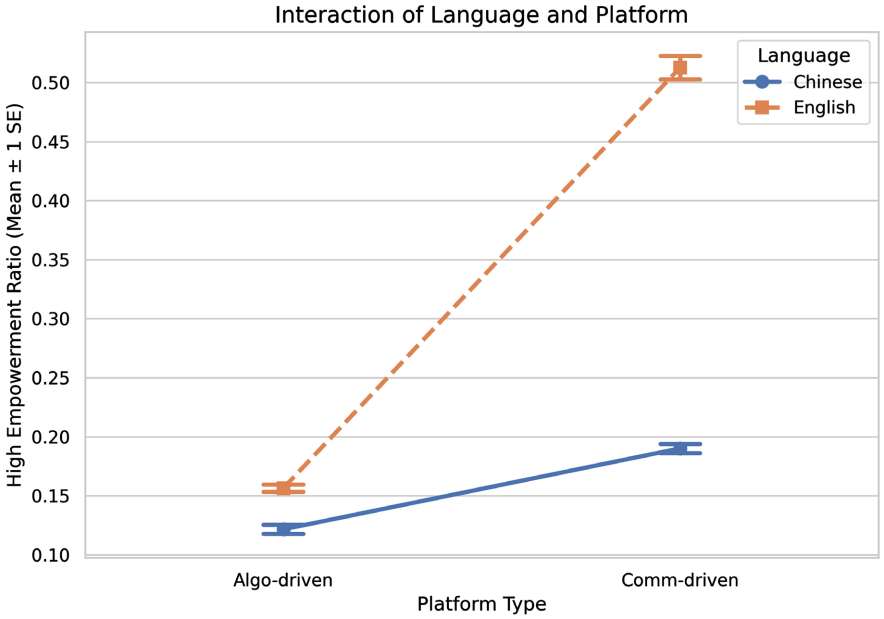


Figure 8. Language and platform interaction effect diagram.

5.3. Theoretical Extension and Supplement

This study improves the theoretical framework through dialogue with existing literature in three aspects: First, it expands the cross-linguistic application scenarios of the digital divide theory, confirms the existence of language heterogeneity in the second and third divides, and fills the theoretical gap; second, deepens the integration of the two major theories, reveals the interactive tension that high affordance does not equal high richness, and provides a new perspective for the application of traditional theories in the digital age; third, extends the social science boundaries of the FinBERT model, combines it with the financial digital divide theory, quantifies financial literacy and empowerment effects, and improves the methodology.

At the level of NLP technology implementation, this study overcomes the multiple challenges of FinBERT cross-language fine-tuning, including differences in Chinese and English word segmentation, the impact of unregistered words, cross-language embedding alignment errors, etc., which not only improves measurement accuracy, but also provides empirical parameters for fine-tuning cross-language financial models.

5.4. Research Limitations

Although this study has made breakthroughs in theory and method, it still has three limitations in combination with the previous research design, which points out the direction for subsequent research: First, at the data level, the sample only covers the entire year of 2024, lacking longitudinal tracking, making it difficult to

capture the dynamic evolution characteristics of the financial digital divide; at the same time, the data only comes from 10 mainstream platforms, and does not cover niche financial communities and regional platforms. There are certain limitations in sample representativeness, which is directly related to the definition of the data collection scope in Section 3.3 above. In addition, although the use of 50 tokens standardization improves cross-language comparability, it still cannot completely eliminate the inherent differences in Chinese and English language structures (such as no inflection in Chinese and different subword splitting logic in English), which may have a slight impact on the absolute comparison of indicators such as term coverage. Subsequent optimization can be further combined with semantic embedding vectors.

Second, at the method level, although cross-language embedding alignment is achieved through unified fine-tuning and the MUSE tool, there are still potential deficiencies in the absolute comparability of the Chinese and English FinBERT models, and the quantification of semantic ambiguity may be affected by language and cultural differences. In addition, although propensity score matching controls some interference variables, it still cannot completely eliminate endogeneity problems. The credibility of causal inference needs to be further improved, and there is a certain subjectivity in the semantic annotation of some texts, which may affect the analysis accuracy. This is also a potential limitation mentioned in the research method section of Section 3.4.

Third, at the variable level, individual user characteristics (such as education level and financial experience) are not included in this study, making it difficult for this study to reveal the moderating effect of individual heterogeneity on the financial digital divide. At the same time, although the definition of high-enabling content refers to the existing framework, it still has a certain degree of subjectivity and may affect measurement accuracy. This limitation also provides a direction for the improvement of the variable system in subsequent research.

6. Conclusions and Future Research Directions

6.1. Research Conclusions

This study is supported by the digital divide theory, media richness theory and platform affordance perspective. Based on the research scope and data collection plan defined in Section 3.3 above, this study selects 35,000 financial text data from 10 Chinese and English social media platforms. It uses the unified and fine-tuned bilingual FinBERT model to combine statistical testing and visual analysis methods to systematically test the impact of language type and platform type on the financial digital divide and the interactive effect of the two. Echoing the three major research hypotheses proposed above, the following core conclusions are drawn:

This study reveals the logic of generating the quality of financial science popularization in the social media environment through multi-dimensional empirical analysis of 35,000 pieces of data. In order to systematically present the research findings, this study constructed **Figure 9** (social media financial information qual-

ity impact mechanism model).

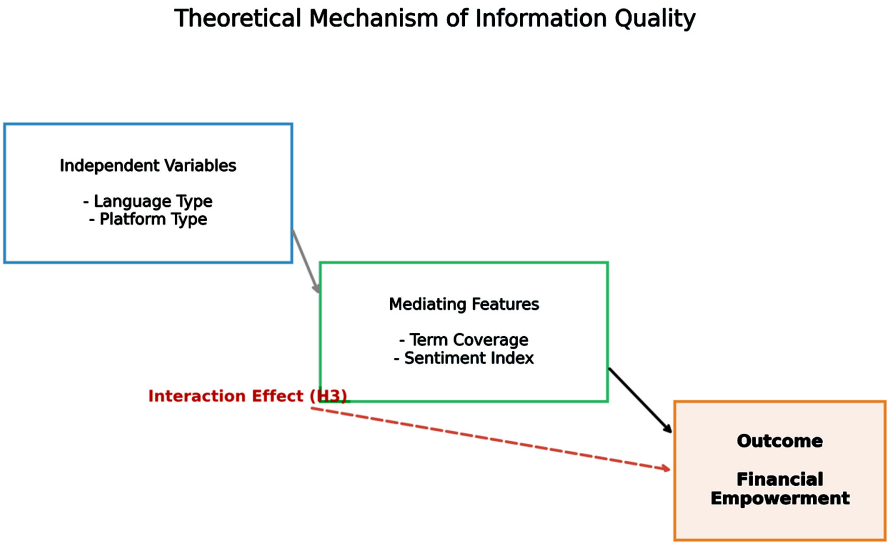


Figure 9. Social media financial information quality impact mechanism model.

Summary of conclusion: This research model reveals that financial information quality is driven by both "language barriers" and "platform logic". The current situation of low empowerment of Chinese financial information is not irreversible and can be alleviated by optimizing the platform recommendation logic and introducing a standardized terminology system. The core conclusions are as follows: First, there is a significant financial expression ability gap in cross-language contexts. The structural differences in the Chinese and English semantic expression systems show that Chinese users have lower coverage of financial terms and higher semantic uncertainty. After controlling for observable characteristics such as text length, Chinese and English users show systematic differences in financial expression ability. This difference is highly consistent with the theoretical expectations of the second financial digital divide. Second, differences in platform affordances create an information quality gap. Algorithmic platforms are more likely to disseminate highly emotional and low-professional content, while community-led platforms are more conducive to the dissemination of professional content. This difference exists across languages and will exacerbate the second gap. Third, the interaction between language and platform significantly affects the cognitive empowerment effect. Language itself is an independent factor in the third gap. Community-led platforms can alleviate the cross-language empowerment gap, while algorithmic platforms amplify the gap, completing the "language-platform" two-dimensional analysis framework.

6.2. Practical Implications

Platform operators need to optimize algorithms, introduce collaborative annotation and professional review mechanisms, increase the weight of standardized financial terminology content, reduce emotional content push, and increase in-

depth content recommendations; community-led platforms strengthen interaction and content screening, and optimize term semantic calibration across language platforms.

Financial science popularization workers should build a systematic Chinese financial terminology system to reduce semantic ambiguity, create science popularization content that combines fragmentation and systematization, implement differentiated science popularization for different platforms, and strengthen terminology application scenario training.

Policymakers need to incorporate the cross-language financial digital divide into inclusive financial policies, increase support for Chinese financial science popularization, standardize platform content distribution, promote the standardization of cross-border financial information services, and formulate differentiated regulatory policies.

6.3. Future Prospects

In the future, research can be deepened from three aspects: first, optimize data and methods, use longitudinal tracking data, expand sample coverage, and combine causal inference methods to reduce endogeneity; second, expand theories and perspectives, explore the moderating role of individual user characteristics, integrate cross-cultural communication theory, and verify the cross-language universality of conclusions; third, extend practical applications, develop financial literacy improvement tools and algorithm optimization plans, conduct cross-regional comparative studies, and combine large language model technology to develop cross-language financial content conversion tools to narrow the cross-language financial digital divide.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Gupta, S. and Chen, H. (2020) Social Media and Financial Information Ecosystem: A Systematic Review. *Journal of Business Ethics*, **165**, 457-478.

[2] Chu, Y., Li, M. and Wang, H. (2022) A Comparative Study of Financial Information Sharing on Chinese and English Social Media Platforms. *Journal of Computer-Mediated Communication*, **27**, 892-910.

[3] Van Dijk, J. (2006) The Deepening Divide: Inequality in the Information Society. Sage Publications. https://doi.org/10.4135/9781452229812

[4] Yang, C., Zhang, Y. and Liu, J. (2020) Cross-Lingual Adaptability of FinBERT Model in Financial Text Analysis. *IEEE Transactions on Knowledge and Data Engineering*, **34**, 3892-3905.

[5] Husin, N. (2024) Social Media Content and Youth Financial Decision-Making: A Quantitative Analysis. *Journal of Youth Studies*, **28**, 213-235.

[6] Hofstede, G. (2011) Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, **2**, 1-26.

https://doi.org/10.9707/2307-0919.1014

[7] Stolper, O.A. and Walter, A. (2017) Financial Literacy, Financial Advice, and Financial Behavior. *Journal of Business Economics*, **87**, 581-643. https://doi.org/10.1007/s11573-017-0853-9

[8] Vassilakopoulou, A. and Hustad, J. (2023) Algorithmic Bias and Financial Digital Divide in the Social Media Era. *New Media & Society*, **25**, 1456-1478.

[9] Li, Z. and Wang, L. (2023) Integrating NLP Technology with Social Science Theories: A Case Study of Financial Digital Inequality Research. *Social Science Computer Review*, **41**, 789-806.

[10] Daft, R.L. and Lengel, R.H. (1986) Organizational Information Requirements, Media Richness and Structural Design. *Management Science*, **32**, 554-571. https://doi.org/10.1287/mnsc.32.5.554

[11] Kim, J., Lee, S. and Park, H. (2021) Cross-Platform Differences in Information Dissemination: A Perspective of Media Richness Theory. *Journal of Communication*, **71**, 289-312.

[12] Angelica, M., Rossi, F. and Verdi, C. (2023) The Impact of Social Media Financial Content on Financial Literacy Enhancement. *Journal of Consumer Affairs*, **57**, 189-210.

[13] Lu, Y., Chen, J. and Zhang, Q. (2023) Internet Infrastructure and Rural Financial Inclusion in China. *China Economic Review*, **79**, Article 101892.

[14] Lusardi, A. and Mitchell, O.S. (2014) Financial Literacy and Planning: Implications for Retirement Wellbeing. *Journal of Economic Perspectives*, **28**, 43-60.

[15] Aissaoui, M. (2022) Objective Measurement of Financial Literacy Using Digital Behavioral Data. *Journal of Financial Counseling and Planning*, **33**, 156-172.

[16] Bucher, T. and Helmond, A. (2018) The Work of Platforms: Reflexivity and the New Media Event. *Information, Communication & Society*, **21**, 22-39.

[17] Burke, K. and Hung, M. (2021) User Engagement in Online Learning Communities: Integrating Media Richness and Affordance Theory. *Computers & Education*, **175**, Article 104389.

[18] Kim, H., Park, J. and Lee, J. (2021) Research Gaps in Cross-Platform Information Behavior Studies. *Library & Information Science Research*, **43**, Article 101032.

[19] Huang, Y., Wang, S. and Li, C. (2023) A Multidimensional Evaluation Framework for Financial Information Quality on Social Media. *Journal of Management Information Systems*, **40**, 567-598.

[20] Amaral, G. and Kolsarici, C. (2020) A Meta-Analysis of Financial Literacy Measurement Methods. *Journal of Economic Surveys*, **34**, 890-912.

[21] Gordon, M.L., Lam, M.S., Park, J.S., Patel, K., Hancock, J., Hashimoto, T., *et al.* (2022) Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In: *Proceedings of the* 2022 *CHI Conference on Human Factors in Computing Systems* (*CHI*'22), Article No.115, 1-19. https://doi.org/10.1145/3491102.3502004

[22] Araci, D. (2019) FinBERT: Financial Sentiment Analysis with Pre-Trained Language Models. http://arxiv.org/abs/1908.10063

[23] Huang, A.H., Wang, H. and Yang, Y. (2023) FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*, **40**, 806-841. https://doi.org/10.1111/1911-3846.12832

[24] Liu, Y., Chen, X. and Wang, Z. (2020) Comparative Analysis of FinBERT and General BERT in Financial Text Processing. *Journal of Financial Data Science*, **2**, 78-92.

[25] Fjellstrom, M. (2022) Stock Market Volatility Prediction Using FinBERT and LSTM Model. *Journal of Forecasting*, **41**, 987-1002.

[26] Chin, T., Lee, K. and Ng, W. (2024) Extracting Investment Signals from Earnings Conference Calls Using FinBERT. *Accounting Horizons*, **38**, 45-62.

[27] Saunders, M., Lewis, P. and Thornhill, A. (2019) Research Methods for Business Students. 8th Edition, Pearson Education Limited.

[28] Hair, J.F., Black, W.C., Babin, B.J. and Anderson, R.E. (2014) Multivariate Data Analysis. 8th Edition, Pearson Prentice Hall.

[29] Hansen, S., McMahon, M. and Prat, A. (2018) Transparency and Deliberation within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics*, **133**, 801-870. https://doi.org/10.1093/qje/qjx045

[30] Pan, L., *et al.* (2020) Cross-Cultural Differences in Financial Discourse: A Corpus-Based Study. *Journal of Pragmatics*, **155**, 120-135.

Appendix A: Examples of Highly Empowering/Non-Empowering Text Annotation

Table A1. Comparison table of examples of highly empowered and non-empowered text annotations.

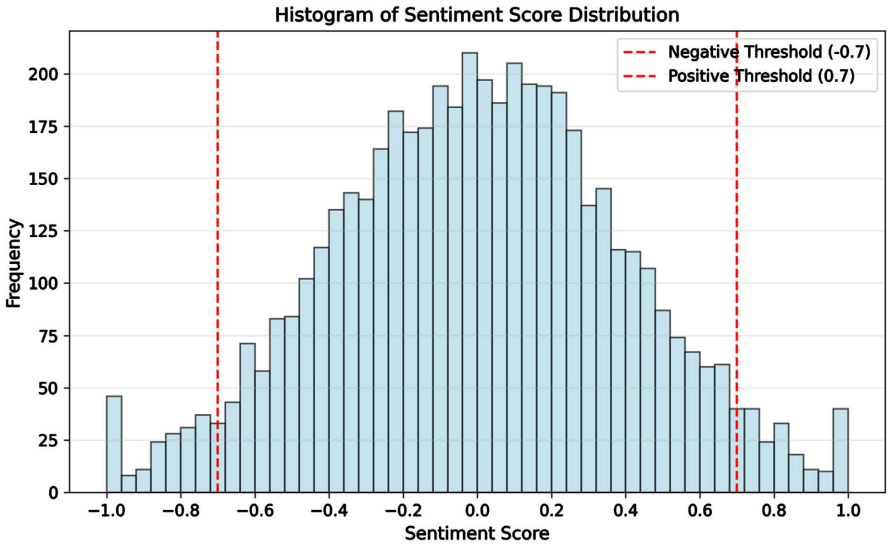
language type	text type	original text	Reason for labeling
Chinese	Highly empowering	The Fed's interest rate hike will increase bond yields, and it is recommended to reduce duration allocation to avoid interest rate risks. The current 10-year government bond yield has exceeded 3.2%, and we need to focus on policy changes in the short term.	It includes mechanism explanation (interest rate increase → rising yield), risk warning (avoiding interest rate risk), data support (10-year government bond yield is 3.2%) and decision-making reference (reducing duration allocation), which fully meets the three empowerment dimensions of "completeness of mechanism explanation, clarity of risk warning, and decision-making reference value".
Chinese	non-empowering	This fund is so profitable, everyone should buy it quickly. If it is too late, you will have no chance!	It contains only emotional recommendations without any financial logic explanation, risk warning or data support, and does not meet the definition of empowering content.
Chinese	Highly empowering	During the A-share annual reporting season, we need to focus on the matching between net profit growth and cash flow. If net profit increases but cash flow from operating activities is negative, there may be a revenue recognition bias. It is recommended to further verify it based on the accounts receivable turnover rate.	It includes interpretation of core indicators (net profit growth, cash flow, accounts receivable turnover rate), risk warning (revenue recognition deviation) and verification methods, and has strong cognitive empowerment value.
Chinese	non-empowering	You are right to follow me and buy stocks. What you bought yesterday has gone up 5 points today!	It only shares personal investment returns, without professional analysis and risk warnings, and is an empirical and emotional expression.
English	Highly empowering	Fed rate hikes increase bond yields; investors should reduce duration to mitigate interest rate risk. The 10-year Treasury yield has exceeded 3.2%, so short-term focus should be on policy changes.	It includes mechanism explanation, risk warning, data support and decision-making suggestions. It meets all three empowerment dimensions and meets the high-enabling content standards.
English	non-empowering	This fund is making a killing—hurry up and buy before it's too late!	It only contains emotional and inflammatory expressions, without professional financial analysis, risk warnings or logical support, and does not constitute empowering content.
English	Highly empowering	During the A-share annual report season, focus on the matching degree between net profit growth and cash flow. Negative operating cash flow with rising net profit may indicate revenue recognition deviations; verify with accounts receivable turnover.	Covers indicator interpretation, risk warning and verification methods, and has complete cognitive empowerment logic.
English	non-empowering	Follow my stock picks and you'll never lose—I bought this yesterday and it's already up 5% today!	It only shares short-term gains, without professional analysis and risk warning, and is an emotional expression based on experience.

1) Qualifications of annotators: The annotation team for this study is a bilingual financial field expert with 10 years of working experience in the information department of China's state-owned banks and 10 years of experience in financial terminology training at Citibank Singapore Branch to ensure the professionalism and consistency of annotation standards. 2) Annotation consistency test: The annotation consistency is verified through Cohen's Kappa test. The Kappa value is 0.82, which meets the annotation quality standard for quantitative research (Kappa ≥ 0.8 is highly consistent). 3) Definition of empowerment dimensions: Highly empowering content must simultaneously meet the three dimensions of "completeness of mechanism explanation", "clarity of risk warning" and "decision-making reference value". If one of them is missing, it will be judged as non-enabling content.

Appendix B: Sentiment Score Distribution and Threshold Calibration

Table B1. Sentiment score percentiles of annotated texts.

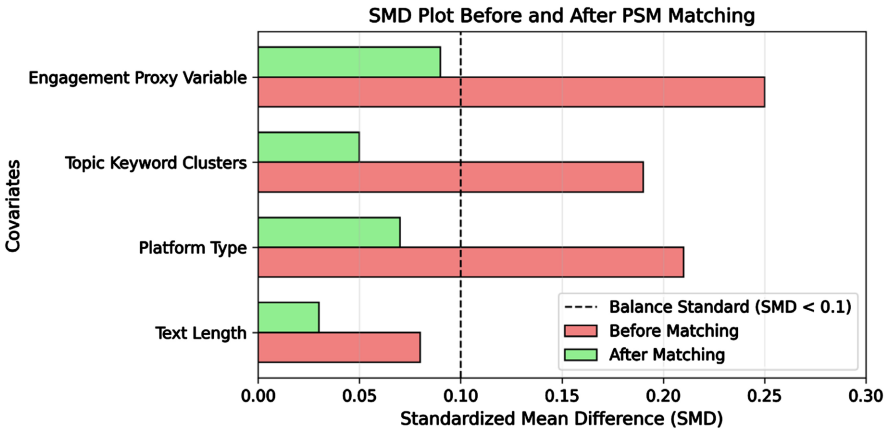
Percentile	Sentiment Score	Interpretation
5th	−0.82	Lower bound of extreme negative sentiment
10th	−0.70	Selected threshold for negative polarization
90th	0.70	Selected threshold for positive polarization
95th	0.83	Upper bound of extreme positive sentiment



Note: Insert histogram with x-axis = sentiment score (−1 - 1), y-axis = frequency; mark thresholds at ±0.7.

Figure B1. Histogram of sentiment score distribution.

Appendix C: PSM Matching Balance Diagnostic Plots



Note: Insert horizontal bar plot comparing SMD values of all covariates before and after matching.

Figure C1. Standardized Mean Difference (SMD) plot before and after matching.