

Factors Influencing User Satisfaction with Artificial Intelligence Applications: An Integrated Approach of Semantic Network and BP Neural Network

-- A Case Study of DouBao APP

Yizhe Wang¹, Chaojie Han², Guan Zhang³, Qiang He^{1,*}

¹ School of Management, Tianjin University of Traditional Chinese Medicine, Tianjin, 301617, China

² Huawei Technologies Co., Ltd., Shanghai, 201206, China

³ Tongfang Knowledge Network Digital Technology Co., Ltd., Beijing, 100083, China

* Corresponding author: Qiang He

Abstract: Generative AI has moved quickly from laboratory demos to everyday mobile apps in China, and user satisfaction has become the clearest test of whether these products actually deliver value. Yet most satisfaction studies still lean on questionnaires — useful, but too structured to capture what people really complain about or praise after prolonged use. This paper looks instead at what users themselves wrote. We collected 63,032 online reviews of ByteDance's DouBao APP (54,248 positive and 8,784 negative, November 2024 to October 2025, compiled via DianDian Data) and asked a simple question: what drives satisfaction, and how much does each driver matter? Rather than relying on a single method, we combined four steps. First, an IWE-AS keyword extraction model — which adds position encoding and a TDF adjustment to the classic TF-IDF — was used to surface terms that genuinely reflect concern, not just frequency. Semantic network mapping (ROSTCM6) and LDA topic modeling (with pyLDAvis for tuning) then revealed how those keywords cluster into themes. Finally, a BP neural network translated those themes into weights in a purely data-driven way, avoiding the subjectivity of expert scoring. The resulting evaluation system covers 6 primary dimensions and 24 secondary items. Interaction (0.347) and accuracy (0.238) dominate — together close to 60% of total weight — followed by intelligence (0.122), practicality (0.116), learning capability (0.101) and overall functionality (0.076). Within those, 'chat / picture generation' and 'failure to understand instructions' in the interaction dimension, and 'speechless / incorrect answers' in the accuracy dimension, carry the highest secondary weights. The implication is practical: for DouBao and similar domestic AI apps, improving how well the system understands colloquial instructions and how reliably it answers — especially for study-related queries — will move satisfaction far more than adding new peripheral features.

Keywords: IWE-AS Model; Semantic Network; LDA Model; BP Neural Network; Online Reviews; User Satisfaction.

1. Introduction

When ChatGPT was released on 30 November 2022, it reached 10 million users in five days. The shock was not just the numbers — it was the proof that a conversational model could behave usefully enough for people to keep coming back. In China, 2023 then became the year of the 'hundred-model competition'. By the latest official count, more than 433 large models have been registered, and the conversation has shifted from 'can we build one?' to 'can we make one people actually want to use?' [1]

What counts as an 'AI application' here matters. We are not talking about a model in a research paper, but software that uses generative AI to simulate human-like help — answering questions, drafting text, creating images or video, tutoring — rather than simply executing fixed rules. The policy backdrop is explicit. As noted in recent central guidance, AI is described as a 'lead goose' technology for this round of industrial transformation, and accelerating its next generation is framed as a strategic choice for capturing the current wave of change [2].

The market has followed. Domestic AI apps on mobile now exceed 270 million monthly active users in aggregate. Within that crowd, DouBao — ByteDance's flagship conversational assistant — has become hard to ignore. By Q3 2025 it

reported about 172 million MAU, ranking first among home-grown AI apps. That scale makes it a useful lens: large enough to generate diverse, real-world feedback, and competitive enough that satisfaction genuinely affects retention. Yet as the market expands, so do expectations. Users no longer judge these apps on whether they 'work' at all, but on how well they understand, how accurate they are, and whether they feel helpful in everyday life.

The usual way to study satisfaction — questionnaires and interviews — struggles here. Response sets are small, prompts are leading, and what people say in a survey often differs from what they type at 11 p.m. when an answer is wrong. Online reviews offer a complement: they are spontaneous, often blunt, and they cover functions, emotions and suggestions in the same paragraph. Compared with a Likert scale, a review that says 'it keeps misunderstanding my spoken instruction and I get angry' tells us not only that interaction failed, but how it felt. This study therefore takes DouBao's mobile reviews as its material and uses text mining plus sentiment analysis to answer two questions: (1) what latent factors shape satisfaction with this kind of AI product? and (2) how much does each factor weigh in practice?

2. Literature Review

2.1. How Satisfaction has been Studied with Review Text

Satisfaction, in the classic sense, is the gap between what a user expected and what they felt they got — pleasure on one side, disappointment on the other [3]. It predicts retention and word-of-mouth, so identifying its drivers accurately matters for product decisions. In the review-mining literature, the playbook is fairly established, but the emphasis differs by domain.

Geng and colleagues, for instance, used JD.com reviews of Apple phones to combine sentiment analysis, semantic networks, word-frequency counts and LDA [4]. Yan et al. worked on a very different service — charging stations in Qingdao — and showed how K-means clustering on keywords followed by a BP network can turn themes into quantified influence scores, with SnowNLP converting text into 5-point satisfaction labels [5]. Chen and Wang stayed closer to product attributes, using a ‘frequency → LDA → regression weighting’ chain to isolate contributions of things like ‘appearance design’ versus ‘logistics speed’ in phone reviews [6]. Feng and Zhao took the health-community route, pairing BERTopic with LSTM sentiment to build a multi-dimensional system for online medical forums [7]. The pattern is consistent: extract, cluster, then weigh. Our work follows that logic, but applies it where it has been less tested — a generative AI assistant rather than a phone or a physical service.

2.2. Keyword Extraction: From TF-IDF to IWE-AS

Methodologically, keyword extraction has seen the most movement. Guo et al. pointed out a simple flaw in classic TF-IDF: it treats all words and positions equally. By adding chi-square statistics for inter-class separation and a position factor that up-weights core terms, they improved feature quality for classification [8]. Sun and Yang proposed IWE-AS, which couples improved word embeddings with an adaptive segmentation strategy to better handle noisy, short user comments [9]. Maisimujiang and co-authors pushed further by stacking BERT, LDA and TextRank — preprocessing, training vectors, mining topics, then re-ranking by similarity — specifically for review analysis [10].

Two gaps motivate what we do next. First, most review studies still target tangible goods or traditional services; generative AI apps — where the ‘product’ is largely a conversation — are under-represented. Second, many AI-satisfaction papers stop at naming dimensions (‘interaction matters’) without saying how much each dimension matters. Without weights, a product team cannot prioritise. We address both by adopting IWE-AS (for its handling of position and TDF-based filtering in messy review text) and by using a BP neural network to learn weights directly from rating behaviour, rather than from expert judgments. The combination is intended to be more faithful to what users actually emphasise when they write.

3. Research Design

3.1. Data and Preprocessing

All reviews come from the DouBao APP mobile platform and were compiled by DianDian Data Network, a third-party mobile analytics provider. The window is November 2024 to

October 2025. After removing duplicates and very short, content-free posts (e.g., ‘good’ alone, emoji-only), we retained 54,248 positive and 8,784 negative texts. That imbalance — roughly 6:1 — is itself informative; DouBao is generally well received, but the negative set is where pain points concentrate. Because the data are unstructured, we built a small pipeline before any modeling: deduplication, short-sentence filtering, a stop-word list, a domain lexicon (to keep terms like ‘DouBao’, ‘search for questions’ and ‘role-play’ intact during segmentation), and then Chinese word segmentation. Only after this cleaning did we count frequencies or train embeddings — otherwise high-frequency noise (‘I’, ‘today’, ‘just’) would dominate the keyword lists.

3.2. How the Analysis Proceeds

The workflow is deliberately staged, and each stage answers a different question:

First, we extract keywords. We favour IWE-AS over raw frequency or standard TF-IDF because, in our early tests, it surfaced more AI-relevant terms (e.g., ‘intelligent’, ‘not understanding’) and suppressed generic praise (‘very good’) through its position and filtering steps.

Second, we map semantics. A semantic network (ROSTCM6) visualises which keywords co-occur and which terms act as hubs — a useful sanity check before we let a topic model decide on the number of themes.

Third, we run LDA to discover latent topics and their high-probability feature words. pyLDavis helps us choose k by inspecting inter-topic overlap; we let positive and negative reviews choose different k values because their vocabularies differ markedly.

Finally, we weigh. The LDA themes plus their keywords become an evaluation system (6 primary, 24 secondary). A BP neural network, fed with theme-level features and star ratings as supervision, learns the non-linear mapping from ‘what was mentioned’ to ‘how satisfied the user was’. The resulting input-to-hidden and hidden-to-output weights give us an objective ranking — no expert panel required. The last step is interpretation: what should actually be fixed first?

4. What the Reviews Actually Talk About

4.1. Keyword Extraction with IWE-AS — Why not Just Count Words?

On messy review text, simple frequency is misleading. A word like ‘good’ appears often but says little about why a user is satisfied; a rarer word like ‘misunderstanding’ appears less but signals a clear failure mode. We compared three routes — raw high-frequency ranking, TF-IDF, and IWE-AS — on the same cleaned corpus. The first two pulled out many generic adjectives; IWE-AS, by contrast, highlighted terms that map more directly to AI-app experience and, crucially, separated what matters in positive versus negative reviews. That is why we kept it.

In practice IWE-AS works in four moves: (1) start from pre-trained Word2Vec vectors for each token — if a word is out of vocabulary, split it into sub-words and take the first match, or, failing that, initialise randomly so the pipeline does not break; (2) compute a TF-IDF-TDF weight, where TDF adds a cross-review distinctiveness signal on top of TF-IDF so that a word that is important in this review but common elsewhere is not over-scored; (3) add a position encoding so that words appearing early (often the main point) or repeated

at the end (often the verdict) receive appropriate emphasis; and (4) filter — vectors whose TF-IDF-TDF falls below 0.2 are zeroed out to cut noise. The remaining weighted vectors

are aggregated in Python and ranked. The top-10 results for positive and negative sets are shown in Tables 1 and 2.

Table 1. Top 10 Keywords and Weights in Positive Reviews

Keywords	Like	Useful	Fine	Good	Intelligent	Recommend	Chat	Great	Practical	Convenient
Weight	0.16	0.14	0.10	0.09	0.06	0.06	0.06	0.05	0.05	0.04

Table 2. Top 10 Keywords and Weights in Negative Reviews

Keywords	Misunderstanding	Answer	Picture	Unwanted	Login	Error	Video	Item	Intelligent	Know
Weight	0.22	0.15	0.08	0.07	0.07	0.06	0.06	0.05	0.04	0.04

The contrast is already telling. Positive reviews cluster around approval and utility — ‘like’, ‘useful’, ‘fine’ — but also around capability words such as ‘intelligent’ and ‘chat’. Negative reviews, by contrast, are led by a comprehension failure (‘misunderstand’, 0.22) and answer quality (‘answer’, 0.15), with visual generation (‘picture’, ‘video’) and functional friction (‘login’, ‘error’) not far behind. In other words, when the app works, users describe it in general evaluative language; when it fails, they point to specific breakdowns.

4.2. Semantic Networks: How Keywords Hang Together

A keyword list says what appears; a semantic network shows what appears together and how strongly. Using ROSTCM6, we built separate co-occurrence networks for positive and negative reviews and visualised them as Figures

1 and 2. Think of each figure as four neighbourhoods around a central hub.

In the positive network (Figure 1), the left neighbourhood gathers around ‘chatting’ and ‘fun’, tying closely to ‘friends’, ‘interesting’, ‘games’ and ‘roles’. Users are clearly valuing the social, playful side of conversation — not just information retrieval, but companionship and role-play. At the bottom, ‘learning’ is the hub, with spokes to ‘help’, ‘answer’, ‘homework’, ‘questions’, ‘tutoring’, ‘children’, ‘improve’ and ‘assistant’. This is the study-aid use case, strong in parent-child and self-learning contexts. On the right, ‘intelligence’ anchors ‘convenient’, ‘fast’, ‘creation’, ‘pictures’ and ‘experience’ — users judging the system on speed and on creative tasks like image generation. Up top, ‘function’ connects to ‘comprehensive’, ‘powerful’, ‘voice’, ‘free’ and ‘practical’, where the emphasis is on breadth of features and the appeal of voice interaction without extra cost.

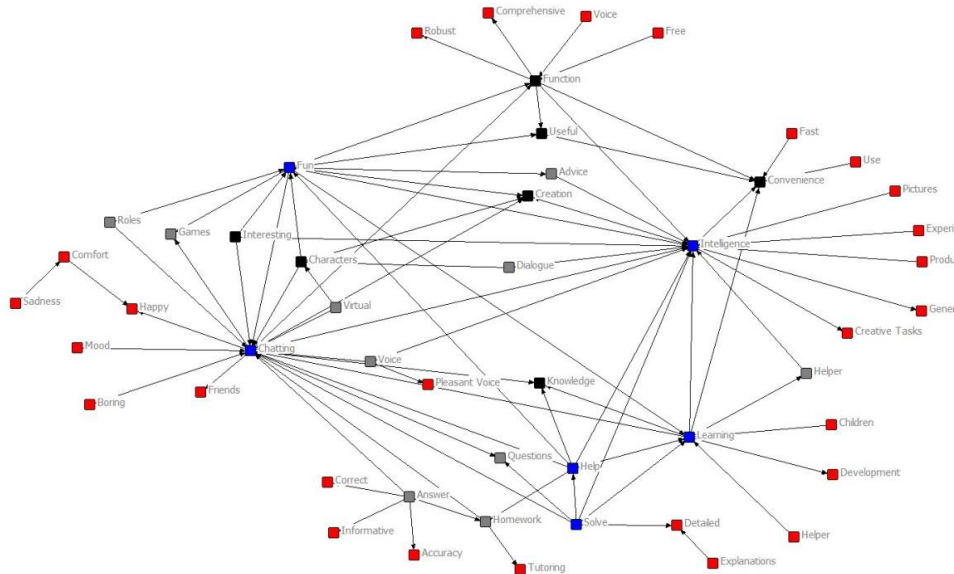


Figure 1. Positive review semantic network diagram

The negative network (Figure 2) tells a different story. The upper zone centres on ‘questions’ and ‘answers’, linked to ‘math problems’, ‘English’, ‘parsing’, ‘accuracy’ and ‘Q&A’ — essentially, ‘can it solve my study question correctly and explain it?’. On the left, ‘errors’ is the hub, with ‘time’, ‘calculation’, ‘recognition’, ‘violation’, ‘homework’, ‘understanding’, ‘ability’ and ‘demand’ around it. Here users catalogue operational failures: slow responses, calculation mistakes, recognition blocks and homework processing faults. On the right, the most emotionally charged hub is ‘not understanding’, tied to ‘instructions’, ‘human-computer’, ‘speech’, ‘angry’ and ‘suggestions’. This is where misinterpretation of colloquial or spoken prompts spills into

frustration — a classic human-computer interaction gap. At the bottom, ‘intelligence’ reappears but now linked to ‘chatting’, ‘roles’, ‘characters’, ‘dialogues’, ‘creation’ and ‘pictures’, suggesting that even when the intelligent features are invoked, they are judged by how coherent the resulting conversation or creation is.

Taken together, the two maps reinforce the same point: positive talk is diffuse and appreciative, negative talk is specific and diagnostic. That specificity is exactly what a product team needs — and what the next step, topic modeling, formalises.

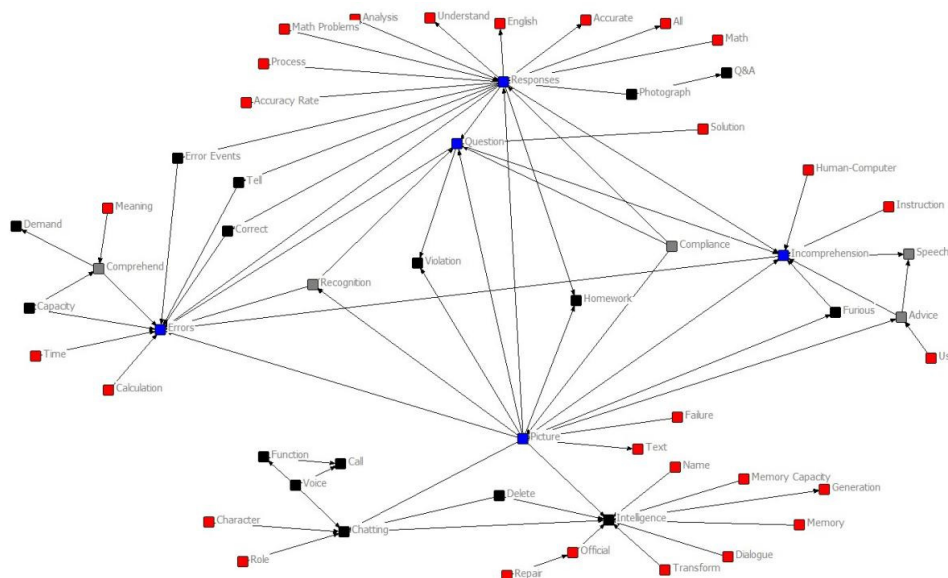


Figure 2. Negative review semantic network diagram

4.3. From Keywords to Topics: LDA Modeling

LDA (Latent Dirichlet Allocation) is unsupervised: it finds latent themes without us labeling them in advance [11]. The crucial choice is k — how many topics to extract. Too few and distinct concerns get lumped together; too many and a single concern shatters into fragments. We used pyLDAvis to

visualise inter-topic distances and overlap for different k values, separately for positive and negative reviews (Figures 3 and 4). For the positive set, $k = 4$ gave the cleanest separation with minimal overlap; for the negative set, $k = 2$ was optimal. Those visual checks matter more than any automatic metric here — if topics overlap heavily in the map, the resulting labels will be hard to interpret.

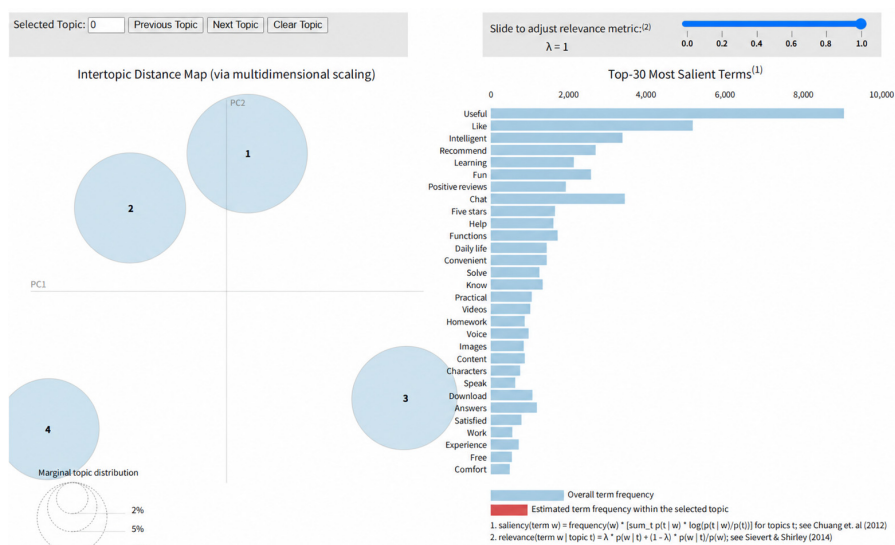


Figure 3. Optimal LDA Topics for Positive Reviews

The outcome was six substantive themes across the two sentiment sets. Rather than treating them as ‘positive topics’ and ‘negative topics’ separately, we merged them into a single evaluation framework — because a theme like ‘intelligence’ appears in both praise and complaint, just with different focus. The six primary dimensions are: practicality, intelligence, learning, function, interaction and accuracy. For each, we retained 14 high-probability feature words (Tables 3 and 4) to ground the subsequent weighting in concrete language users actually used, not in abstract labels we invented.

5. How Much Does Each Factor Matter? Weighting with a BP Neural Network

Naming themes is only half the job. The harder question — and the one product teams actually need answered — is how much each theme moves satisfaction. Expert scoring (e.g., AHP) can do this, but it imports the experts’ biases. We preferred to let the data speak: if mentions of a theme systematically co-vary with star ratings, that theme matters more, regardless of what anyone thinks should matter.

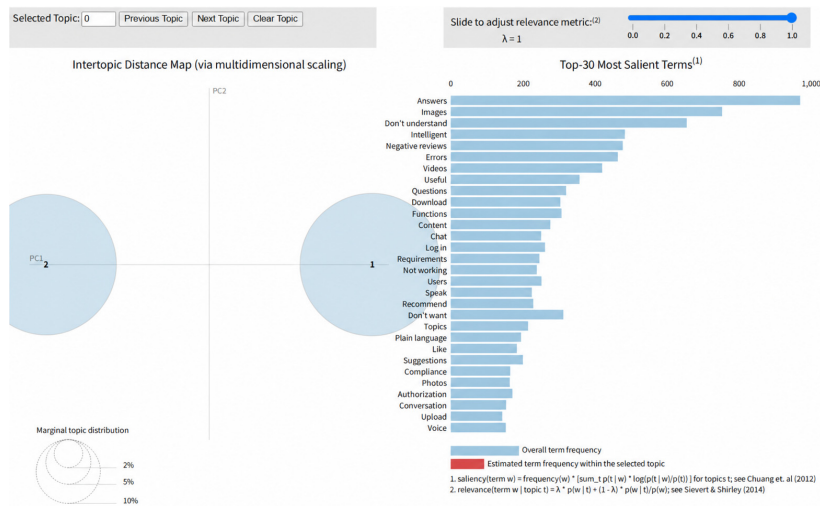


Figure 4. Optimal LDA Topics for Negative Reviews

Table 3. Topics and Keywords of Positive Online Reviews

Number	Topic	High-Probability Feature Words
Topic0	Practicality	Learning, work, life, problem, explanation, like, helpful, useful, good helper, companionship, solve, practical, assistant, knowledge
Topic1	Intelligence	Intelligent, accurate, creation, analysis, thoughtful, reply, satisfied, worthy, experience, know, answer, information, time, comprehensive
Topic2	Learning	Homework, question, English, search for questions, solve problems, useful, download, free, suggestion, answer, child, tell, speak, teacher
Topic3	Function	Video, picture, voice, chat, photo, recommend, happy, fun, value, sound, function, create, dialogue, suggestion

Table 4. Topics and Keywords of Negative Online Reviews

Number	Topic	High-Probability Feature Words
Topic0	Interaction	Fail to understand, experience, chat, upload, display, memory, picture, violation, character, novel, requirement, dialogue, speak, upload
Topic1	Accuracy	Negative review, error, speechless, incorrect, answer, question, search for questions, know, topic, think, understand, meaning, information, search

Table 5. Index Weight Table

First-Level Indicators	Weight	Second-Level Indicators	Weight
Interaction	0.347	Chat	0.992
		Picture	0.992
		Fail to understand	0.496
		Dialogue	0.496
Accuracy	0.238	Speechless	0.380
		Incorrect	0.024
		Search for questions	0.024
		Answer	0.012
Intelligence	0.122	Accurate	0.356
		Comprehensive	0.331
		Intelligent	0.153
		Thoughtful	0.076
Practicality	0.116	Great	0.384
		Practical	0.251
		Learning	0.075
		Helpful	0.050
Learning	0.101	Useful	0.906
		Free	0.031
		Suggestion	0.031
		Key	0.022
Function	0.076	Advice	0.073
		Fun	0.062
		Picture	0.052
		Happy	0.042

A BP (back-propagation) neural network fits this setting well. It is supervised, learns non-linear mappings between

inputs and outputs through iterative training, and, importantly, is purely data-driven in how it assigns importance — a useful property when the corpus is large (63k reviews) and the language is varied and noisy. In our setup, theme-level features are the inputs and satisfaction scores are the outputs.

Concretely, we did the following. First, we kept only reviews with clear polarity — ratings 1, 2, 4 and 5 — and dropped the three neutral comments that sat on the boundary. To handle the imbalance (far more 5-star reviews), we down-sampled the 5-star set, keeping roughly one in five, so the model would not learn to predict ‘everyone is happy’. Second, for each review we matched its text against the six themes via their high-probability keywords, counted matches per theme, and normalised to obtain theme feature values. Third, we built a BP network with theme features as the input layer and satisfaction as the output, normalised the data, split into training and test sets, and trained until convergence. Finally, we extracted the input→hidden and hidden→output weight matrices, summed and normalised them to theme-level influence, and computed keyword importance as (keyword’s share of its theme’s matches) × (theme weight). The full hierarchy is in Table 5.

6. Results and Discussion

Stepping back, what did we actually find? Using DouBao’s 63,032 reviews from November 2024 to October 2025, and moving from keyword extraction (IWE-AS) through semantic networks and LDA to a data-driven weighting (BP network), a consistent picture emerges. Satisfaction with a generative AI assistant is not primarily about how many features it lists on a download page. It is about two things working reliably: whether the system understands what the user meant — especially spoken or colloquial instructions — and whether the answer it gives is correct and well explained.

That is why interaction and accuracy together account for nearly three-fifths of the weight. Users spend a lot of their review text on ‘it didn’t understand me’ moments (misunderstanding weight 0.22 in the negative keyword list; ‘fail to understand’ 0.496 as a secondary indicator) and on wrong or ‘speechless’ answers (0.380). Picture and chat functions are highly weighted not because users love novelty alone, but because those are the contexts where misunderstanding is most visible — a misinterpreted prompt yields a useless image, and the disappointment is immediate. By contrast, broader ‘intelligence’ or ‘practicality’ — useful, thoughtful, great — matter, but more as amplifiers when the core works, not as substitutes when it doesn’t.

For product teams, the prioritisation is therefore fairly concrete. First, invest in instruction understanding — handling of spoken input, colloquial phrasing, and multi-turn context — before adding peripheral features; the negative semantic network’s ‘not understanding → angry → suggestions’ chain shows how quickly comprehension failures erode goodwill. Second, shore up accuracy in the study-assistance scenario (math, English, parsing, solution explanations), which dominates both the ‘questions/answers’ hub and the accuracy theme; users explicitly judge the app on whether explanations are detailed, not just whether the final answer is right. Third, keep what already works: the learning-assistance and creative-generation experiences that drive positive mentions of ‘helpful’, ‘convenient’ and ‘useful’. These are retention strengths, even if they weigh less than the

two risk factors.

We should also be candid about limits. The data are from one app and one year; cross-app comparison would test whether the 60% dominance of interaction + accuracy generalises beyond DouBao. Reviews are also self-selected — people who bother to write are often those who are delighted or frustrated, not the silent middle. And while the BP network gives objective weights, it does so at the cost of interpretability; we report the weights, but the network itself remains a black box. Replicating the pipeline on other domestic AI apps, and complementing it with a small set of interviews to hear the ‘silent majority’, would be natural next steps. What this study adds is not a claim that we have the final satisfaction model, but a reproducible, largely automated way to surface — and rank — what users themselves keep returning to in their own words.

References

- [1] Li, X., & Wang, Y. (2024). The development trend of global large-scale artificial intelligence models and their industrial impact. *Journal of Artificial Intelligence Research*, 62(1), 45–78. <https://doi.org/10.1613/jair.1.15243>.
- [2] Ministry of Science and Technology of the People's Republic of China. (2023). Strategic guidelines for accelerating the development of a new generation of artificial intelligence. *Science & Technology Review*, 41(8), 5–18. <https://doi.org/10.3969/j.issn.1000-7857.2023.08.001>.
- [3] Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research*, 17(4), 460–469. <https://doi.org/10.1177/002224378001700405>.
- [4] Geng, X. L., & Zhang, H. (2024). Exploring user satisfaction factors of e-commerce products based on text mining: A case study of JD Apple mobile phones. *Computational Economics*, 63(3), 1123–1145. <https://doi.org/10.1007/s10614-023-10324-x>.
- [5] Yan, X. Y., Liu, J., & Chen, Z. (2023). Quantifying user satisfaction of charging stations via K-Means clustering and BP neural network. *Sustainable Cities and Society*, 98, 104789. <https://doi.org/10.1016/j.scs.2023.104789>.
- [6] Chen, Z. J., & Wang, Q. (2024). Analyzing the contribution of product attributes to user satisfaction: Evidence from JD mobile phone reviews. *Electronic Commerce Research and Applications*, 58, 101782. <https://doi.org/10.1016/j.elerap.2024.101782>.
- [7] Feng, Y., & Zhao, M. (2023). Constructing a multi-dimensional satisfaction evaluation system for online health communities using BERTopic and LSTM. *Journal of Medical Internet Research*, 25(6), e43215. <https://doi.org/10.2196/43215>.
- [8] Guo, T. Y., & Li, S. (2024). Optimizing TF-IDF for text classification: Integrating chi-square statistics and position weight. *Pattern Recognition Letters*, 178, 109–117. <https://doi.org/10.1016/j.patrec.2024.02.005>.
- [9] Sun, G. Y., & Yang, H. (2023). IWE-AS: A text classification model based on improved word embedding and adaptive segmentation. *IEEE Transactions on Knowledge and Data Engineering*, 35(12), 11890–11902. <https://doi.org/10.1109/TKDE.2023.3289456>.
- [10] Maisimujiang, N., & Abudureyimu, A. (2024). A keyword extraction method integrating BERT, LDA, and TextRank for user review analysis. *Neural Computing and Applications*, 36(8), 6215–6232. <https://doi.org/10.1007/s00521-023-08876-9>.
- [11] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022. <https://doi.org/10.5555/944919.944940>.