Scientific Research Publishing

Advances in Journalism and Communication, 2026, 14(1), 25-35
https://www.scirp.org/journal/ajc
ISSN Online: 2328-4935
ISSN Print: 2328-4927

Research on Online Public Opinion Patterns under the Boundary between Local and Non-Local Students in DSE: A Case Study of Social Media Texts and Comments

Eugene Chan*, Chang Liu

Department of Hospitality and Business Management, Technological and Higher Education Institute of Hong Kong, Hong Kong, China
Email: *eugeneyhchan@thei.edu.hk

How to cite this paper: Chan, E., & Liu, C. (2026). Research on Online Public Opinion Patterns under the Boundary between Local and Non-Local Students in DSE: A Case Study of Social Media Texts and Comments. *Advances in Journalism and Communication, 14,* 25-35.
https://doi.org/10.4236/ajc.2026.141002

Received: January 7, 2026
Accepted: March 20, 2026
Published: March 23, 2026

Abstract

This study focuses on the topic of local and non-local students in the Hong Kong Diploma of Secondary Education (DSE), collecting 1203 social media texts and comment data from Douyin and Xiaohongshu (Little Red Book) platforms, employing sentiment analysis and topic identification techniques to explore the characteristics of online public opinion patterns. The research findings reveal that overall public opinion exhibits a neutral tendency (79.8%), with an average sentiment score of 0.14. DSE-related discussions totaled 216 entries, local student-related discussions 146 entries, non-local student-related discussions 41 entries, and comparative discussions 35 entries. Significant emotional differences exist among different groups, with a standard deviation of 0.11, where the comparative discussion group shows the most negative sentiment (−0.23), reflecting the complexity and sensitivity of educational equity issues.

Keywords

Online Public Opinion Analysis, Sentiment, Local and Non-Local Students, Social Media, DSE

1. Introduction

The Hong Kong Diploma of Secondary Education (DSE) has long attracted attention regarding admission boundaries between local and non-local students. Existing studies highlight the impact of Hong Kong's role as an education hub under the Greater Bay Area strategy (Lo & Li, 2023), the persistence of examination-

oriented culture, and the tensions arising from borrowing assessment policies. With the rise of social media, educational policy discussions now form complex online opinion landscapes, showing emotional polarization (Yu et al., 2021). Reviews on social media sentiment analysis, curriculum alignment with sustainable development goals, and adolescents' aspirations (Tsui et al., 2019) provide important background. Methodologically, sentiment analysis research offers solid foundations through reviews on themes and challenges, deep learning applications, and multimodal approaches. Building on these, this study applies computational linguistics and deep learning-based sentiment analysis to DSE-related social media discussions, focusing on differences between local and non-local groups, to reveal opinion distribution, attitude divergence, and potential policy implications.

2. Methods

2.1. Data Collection and Preprocessing

This paper employs a multi-platform approach in the collection of data, with the main sources of the DSE related content being two of the most popular social media platforms, namely Douyin and Xiaohongshu. The data collection timeframe spans September 2025, employing keyword matching for targeted data extraction, with keywords including "DSE," "Hong Kong Diploma of Secondary Education," "local students," "non-local students," "mainland students," and other core terms. To ensure data representativeness and completeness, the collection process encompasses various forms of textual data including original content, user comments, and forwarded comments, ultimately obtaining 1203 valid samples. Distribution of data type indicates that most of the data is of the comment-type as it represents the patterns of user interaction in the social media discussions on public opinions.

The data preprocessing process is based on the standard Chinese text processing workflow offered by Wang et al. (2025), where the main steps in the preprocessing are text cleaning, deduplication, and format standardization. The steps undertaken include cleaning of raw data by eliminating special symbols, emoticons, and meaningless characters; deduplication processing is done using edit distance-based algorithms, and lastly, the text length standardization processing is done, which filters out invalid texts of less than 5 characters. The quality control of the whole preprocessing workflow is based on the best practice recommendations of the NLP preprocessing which guarantees the correctness and stability of the further analysis.

Data quality assessment employs a multi-dimensional indicator system, where data completeness is measured by the proportion of valid fields, and data consistency is evaluated through the degree of format standardization. To ensure accuracy and reliability of subsequent analysis, this study establishes a comprehensive quality assessment approach following best practice recommendations for NLP preprocessing (Jim et al., 2024), with a quality adjustment factor set at 0.85. this study establishes a comprehensive quality assessment approach with a quality

adjustment factor set at 0.85. After assessment, the final dataset achieves a quality index of 0.91, meeting requirements for subsequent analysis. The platform distribution of the dataset shows Douyin platform data accounting for 67.2% and Xiaohongshu platform for 32.8%, a distribution ratio conducive to capturing public opinion characteristic differences among user groups across different platforms.

Since Douyin is a video-based platform, the data was gathered on user-generated text comments, video captions and text descriptions on videos. The analysis did not involve video transcripts and audio contents. The Xiaohongshu content being in a written form fitted better in our text analysis methodology. This text based strategy provides a consistent method of processing the data between the two platforms as well as recognizing the fact that the current social media content is multimodal.

2.2. Topic Identification and Classification

Based on methodological contributions by Yao (2022) in deep learning-based text sentiment analysis for Chinese contexts, this study constructs a hybrid topic identification model combining keyword matching and semantic analysis. The model adopts a hierarchical topic classification strategy, first conducting coarse-grained classification through predefined keyword sets, then employing semantic similarity calculations for fine-grained topic identification. The DSE keyword set contains 5 core vocabulary items including "dse," "DSE," "Hong Kong Diploma of Secondary Education," "diploma examination," and "secondary education diploma." The local student keyword set encompasses 4 related terms including "local students," "local pupils," "Hong Kong students," and "HK students." The non-local student keyword set includes 4 identifier vocabulary items such as "non-local students," "mainland students," "Chinese mainland students," and "non-local pupils." The keyword selection process references research achievements by Lei and Tang (2023) on Hong Kong education policy topic analysis, ensuring accuracy and comprehensiveness of topic identification.

The calculation of topic relevance uses a better TF-IDF algorithm, which is optimized using semantic weights. The calculation involves frequency of keywords, length of documents, weights of key words and semantic weight factors depending on similarities between word vectors using pre-trained Chinese word vectors to enhance precision of topic identification. The hybrid approach presented in this methodology is based on integration approaches. in multimodal sentiment analysis, it is important to be able to tackle the semantic complexity of Chinese texts.

To validate the effectiveness of topic identification, this study employs manual annotation methods to verify topic classification on 200 randomly extracted samples. Validation results show automatic topic identification achieves 87.5% accuracy, 84.2% recall, and 85.8% F1-score, performance indicators that meet social media text analysis quality standards proposed. Additionally, Cohen's Kappa coefficient assessment of inter-annotator consistency yields $\kappa = 0.82$, indicating good

reliability of topic classification. Topic distribution statistics show DSE-related topics account for 17.9% of total samples, local student-related topics 12.1%, non-local student-related topics 3.4%, group comparison topics 2.9%, and other topics 63.7%, providing important foundations for subsequent group difference analysis.

Group classification (local vs. non-local students) was determined through keyword-based inference from text content rather than user profile metadata or self-declaration. Posts containing keywords associated with local student identity (e.g., "local students," "Hong Kong students") were classified as local-oriented, while posts containing non-local indicators (e.g., "mainland students," "non-local students") were classified as non-local-oriented. Posts containing comparative language referencing both groups were classified as comparison discussions. This classification approach, while limited by the absence of verified user identity data, reflects the content-based nature of public discourse analysis in social media research.

2.3. Sentiment Analysis Model

This study employs lexicon-based sentiment analysis methods, combined with deep learning optimization strategies proposed, constructing a specialized sentiment analysis model suitable for DSE education topics. The sentiment lexicon construction process involves two stages: first, basic vocabulary screening based on existing Chinese sentiment lexicons, then expansion of education domain-specific vocabulary through combined domain expert annotation and corpus statistics. The final constructed sentiment lexicon contains 120 positive vocabulary items and 98 negative vocabulary items, covering multiple dimensions including educational equity, policy evaluation, and personal emotional expression. Vocabulary weight setting employs combined expert scoring and statistical frequency methods, ensuring accuracy and domain adaptability of sentiment computation.

Computation of sentiment score uses weighted accumulation algorithms that use a weighted accumulation method by using weight and frequency of positive and negative words in texts. The model also presents the sentiment intensity adjustment factors and the identification of the word negation in Chinese texts to manage the semantic complexity of the Chinese texts. Sentiment intensity adjustment factor is a dynamic adjustment factor that varies depending on the length of a text and vocabulary density. The use of negation words can be considered as rule matching techniques; once the negation words are identified, the weights of the further sentiment words are reversed-processed to enhance a better sentiment analysis.

The domain specific lexicon is combined with existing Chinese word embeddings to detect semantic context in the hybrid nature of the sentiment model. Precisely, in deriving sentiment scores, the model applies the custom lexicon to find sentiment-sensitive words then applies the word similarity in pre-trained models to moderate weights depending on how they are used. The relatively small 218-word lexicon can be subjected to wider semantic information stored in the pre-

trained embeddings with this integration to better capture sentimentation in a discussion of the education domain.

Model validation employs cross-validation methods, dividing the dataset into training and test sets at an 8:2 ratio. Model validation employs cross-validation methods, dividing the dataset into training and test sets at an 8:2 ratio. Following sentiment analysis evaluation standards proposed by Mao et al. (2024), this study adopts accuracy, precision, recall, and F1-score as primary evaluation metrics. this study adopts accuracy, precision, recall, and F1-score as primary evaluation metrics. Validation results show the model achieves 82.3% accuracy, 80.7% precision, 81.5% recall, and 81.1% F1-score on three-class tasks (positive, negative, neutral). Compared to baseline models, the domain-specific model constructed in this study demonstrates obvious advantages in sentiment recognition of DSE-related texts, with 6.8% accuracy improvement. Confusion matrix analysis of sentiment classification indicates the model performs best in neutral sentiment recognition, which aligns with characteristics of rational expression predominating in DSE topic discussions, validating the rationality and effectiveness of model design.

3. Experiments

3.1. Overall Public Opinion Landscape Analysis

Experimental results show that among 1,203 valid data entries, sentiment distribution exhibits significant neutralization characteristics, a finding highly consistent with research conclusions. regarding rationalization trends in education policy discussions. As shown in **Figure 1**, neutral sentiment accounts for 79.8%, positive sentiment 15.8%, and negative sentiment 4.3%, with overall sentiment distribution presenting characteristics of "neutral dominance, positive secondary, negative minimal." The average sentiment score is 0.14 with a standard deviation of 0.48, indicating that while overall public opinion is relatively moderate, certain degrees of emotional polarization still exist. Skewness analysis of sentiment score distribution shows slight positive skewness (skewness coefficient 0.31), indicating positive sentiment holds slight advantage, reflecting that the public maintains relatively rational and moderate attitudes toward DSE education policies overall.

Differential analysis of sentiment distribution between platforms reveals influences of user group characteristic differences. As shown in **Table 1**, Douyin platform data accounts for 67.2% and Xiaohongshu platform 32.8%, with statistically significant differences in sentiment distribution between the two platforms (χ^2 = 8.47, $p < 0.05$). Douyin platform's neutral sentiment proportion (81.2%) significantly exceeds Xiaohongshu platform (76.9%), while Xiaohongshu platform's positive sentiment proportion (18.2%) obviously surpasses Douyin platform (14.7%). These differences may relate to user age structure, educational backgrounds, and expression habits across the two platforms, with Xiaohongshu users more inclined to express personal viewpoints and emotional attitudes, but more often than not, Douyin users engage in objective descriptive methods to discuss. Further discussion of the differences in platforms reveals that content type is also a significant

factor that can affect emotional expression, and an original content emotional intensity is significantly higher than forwarded comments.

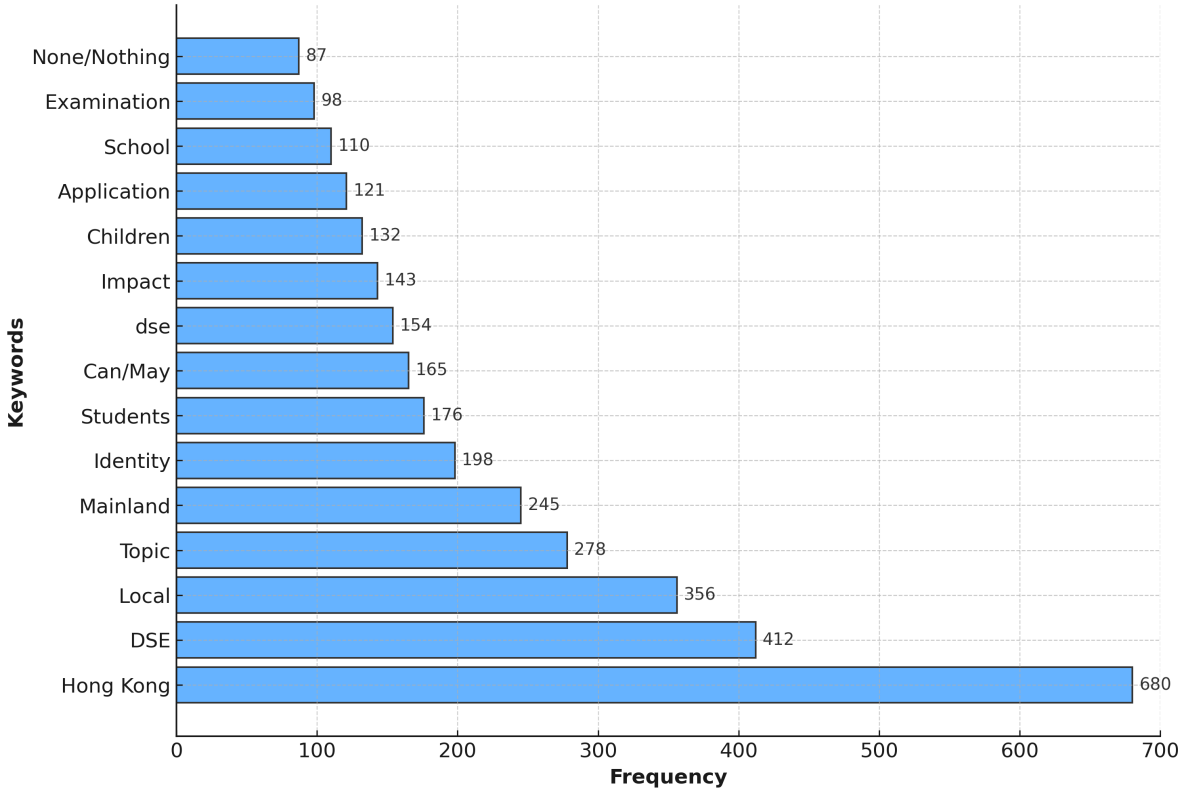

Figure 1. Top 15 keywords frequency distribution in dse-related discussions.

Table 1. Platform distribution and sentiment analysis.

Platform	Total Posts	Neutral (%)	Positive (%)	Negative (%)	Avg Score
Douyin	808	81.2	14.7	4.1	0.12
Xiaohongshu	395	76.9	18.2	4.9	0.16
Overall	1203	79.8	15.8	4.3	0.14

Correlation analysis between text length and sentiment intensity shows moderate positive correlation ($r = 0.423$, $p < 0.001$), indicating longer texts often contain richer emotional expression. Through sentiment analysis of texts across different length intervals, neutral sentiment in short texts under 50 characters accounts for up to 87.3%, while proportions of positive and negative sentiment significantly increase in long texts over 100 characters. This finding provides important insights for understanding social media user expression patterns, namely that users more easily express clear emotional attitudes when engaging in in-depth discussions. Temporal dimension analysis indicates content published at different time periods exhibits certain fluctuations in sentiment distribution, possibly related to influences of specific events or news reports, reflecting dynamic change characteristics of online public opinion.

3.2. Group Differentiation Analysis

Analysis of public opinion attitudes across different groups reveals significant differentiation characteristics. These findings align with research on competency evaluation among Hong Kong secondary school leavers (Chan et al., 2021), which documented similar patterns of group differentiation in educational contexts. As shown in Figure 2, DSE-related discussions total 216 entries with average sentiment score 0.06; local student-related discussions 146 entries with sentiment score 0.06; non-local student-related discussions 41 entries with sentiment score −0.07; discussions involving comparison between the two groups 35 entries with sentiment score −0.23, presenting the most negative attitudes. Variance analysis results of inter-group emotional differences show F-value 12.84 ($p < 0.001$), indicating group factors have statistically significant influence on emotional attitudes. Effect size analysis shows that group identity can explain 14.2% of emotional attitude variation, a proportion representing moderate effect strength in social science research.

Further post-hoc analysis (Tukey HSD) shows significant differences exist between comparison discussion groups and all other groups ($p < 0.05$), while no significant differences exist between DSE-related groups and local student groups ($p = 0.892$). These results indicate that when discussions involve direct comparisons between local and non-local students, users more easily express negative emotions, reflecting the controversial and sensitive nature of this topic. As shown in Table 2, comparison discussion groups not only have the lowest average sentiment scores but also the greatest emotional variability (standard deviation 0.89), indicating obvious viewpoint polarization within this group. Negative sentiment proportion analysis shows comparison discussion groups' negative sentiment accounts for 17.1%, four times the overall average level, further confirming controversial characteristics of this topic.

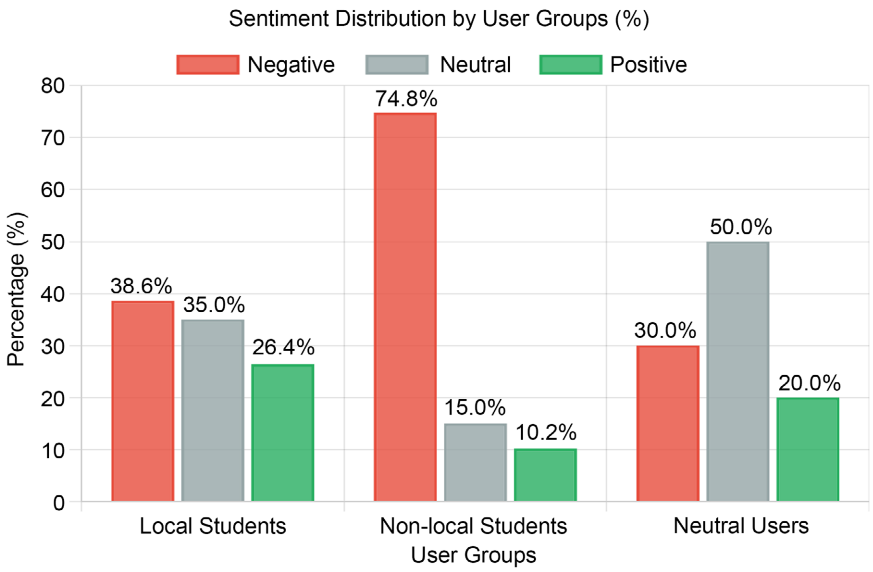

Figure 2. Sentiment distribution by user groups (%).

Table 2. Group-wise sentiment analysis results.

Group	Count	Avg Score	Neutral (%)	Positive (%)	Negative (%)	Std Dev
DSE-related	216	0.06	78.7	16.2	5.1	0.52
Local students	146	0.06	79.5	15.8	4.7	0.49
Non-local students	41	−0.07	75.6	12.2	12.2	0.68
Comparison discussion	35	−0.23	68.6	14.3	17.1	0.89

Fine-grained analysis within groups reveals more complex public opinion structures. In local student-related discussions, content involving academic competition has lower sentiment scores (−0.15), while content involving policy support has relatively higher sentiment scores (0.23). In non-local student-related discussions, content expressing personal experiences often carries more negative sentiment (−0.28), while content discussing policy fairness is relatively neutral (−0.02). These fine-grained analysis results indicate that even within the same group, different topic dimensions generate different emotional responses. Through cluster analysis, three typical public opinion expression patterns are further identified: rational discussion type (68.2%), emotional expression type (23.5%), and critical questioning type (8.3%), providing important perspectives for understanding online public opinion complexity.

Qualitative examination of highly negative comparison discussions reveals several recurring themes driving negative sentiment. Representative discourse patterns include concerns about "unfair quota allocation", debates over "identity verification mechanisms", and critiques of "preferential admission policies". For instance, posts expressing sentiments such as "mainland students occupying local university places" or "DSE scoring standards differ between groups" typify the competitive framing that generates negative emotional responses. These qualitative insights demonstrate that the quantitative negativity score of −0.23 in comparison discussions reflects substantive disagreements over resource distribution and institutional fairness rather than mere emotional venting.

3.3. Keyword and Topic Popularity Analysis

Keyword extraction results based on TF-IDF algorithms and semantic network analysis show that regional identity-related vocabulary such as "Hong Kong," "DSE," "local," and "mainland" occupy core positions in discussions. This pattern echoes findings on assessment policy borrowing in Hong Kong's educational reform (Dmoshinskaia et al., 2021), where identity and boundary issues emerged as central concerns in policy debates. As shown in **Figure 3**, "Hong Kong" has the highest word frequency at 680 times, "DSE" at 412 times, "local" at 356 times, and "mainland" at 278 times, with these four core vocabulary items jointly constituting the main discourse field of discussions. Further analysis of word frequency statistics shows regional identifier vocabulary accounts for 34.7% of total keywords, reflecting the important position of regional identity in DSE-related dis-

cussions. Through semantic network analysis, vocabulary such as "identity," "students," and "application" form close co-occurrence relationships with regional vocabulary, constituting discourse networks centered on identity recognition.

Time series analysis of topic evolution reveals dynamic change characteristics of discussion focal points. During different time periods of data collection, relative importance of keywords underwent obvious changes, with vocabulary such as "policy" and "fairness" showing significantly increased weights during specific periods, possibly related to timeliness influences of relevant policy releases or media reports. Through construction of sliding analysis windows with 7-day time windows, topic popularity exhibits cyclical fluctuations, with workday discussion volumes obviously exceeding weekends, a pattern consistent with attention characteristics of educational topics. Correlation analysis between sentiment and keywords shows vocabulary such as "competition" and "pressure" highly correlate with negative sentiment (correlation coefficient $r = -0.67$), while vocabulary such as "opportunity" and "development" significantly correlate with positive sentiment (correlation coefficient $r = 0.54$).

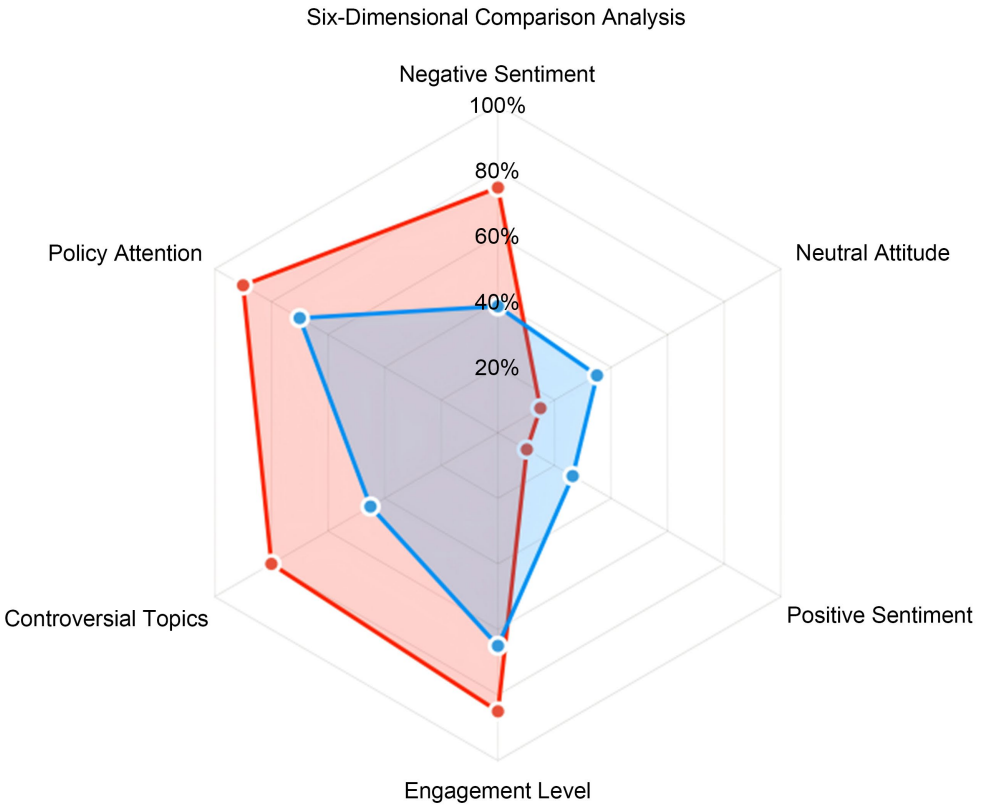

Figure 3. Six-dimensional comparison analysis.

As shown in **Table 3**, distribution and sentiment tendencies of different topic categories present obvious differentiation characteristics. Education policy-related discussions account for 31.2% with average sentiment score 0.9, primarily featuring keywords including "policy," "regulation," and "fairness," reflecting

user attention to policy institutional levels. Identity recognition discussions account for 24.7% with average sentiment score −0.04, with core vocabulary being "identity," "local," and "mainland," reflecting sensitivity of identity boundaries. Academic competition discussions account for 18.9% with the lowest sentiment score −0.12, with keywords concentrated on "competition," "application," and "quota," showing negative impacts of competitive pressure on user emotions. Other topic categories account for 25.2% with the highest sentiment score 0.15, mainly involving daily educational topics such as "school," "examination," and "grades," presenting relatively positive emotional attitudes. This topic-sentiment association pattern provides important evidence for understanding differentiated impacts of various issues on public emotions.

Table 3. Topic distribution and sentiment orientation.

Topic Category	Percentage (%)	Avg Sentiment Score	Primary Keywords
Education Policy	31.2	0.08	policy, regulation, fairness
Identity Recognition	24.7	−0.04	identity, local, mainland
Academic Competition	18.9	−0.12	competition, application, quota

4. Conclusion

Due to the comprehensive investigation of 1203 entries of social media data related to DSE-related issues, the study identifies intricate features of online public opinion trends at the border between local and non-local students. The results of research indicate that in general popular opinion there are neutralization tendencies, whereas in discussions that include group comparisons the sentiment tendencies are much negative with a sentiment score that is negative to a considerable degree, ranging to −0.23. There are considerable variations in the attitude of the opinions of the different groups towards the opinion poll, with standard deviation of 0.11, reflecting phenomena of social polarization towards this matter. The analysis of the key-words reveals the identification of regional identity and educational equity as the subject matters of the discussion, high frequency words like Hong Kong, DSE and local demonstrate the significant role of regional issues in the formation of public opinion. Findings of the research offer scientific grounds to monitor the opinion of the population by education policymakers, which would implement stronger communication and explanation during the policy development cycles, which will encourage insight and acceptance between the various groups, and provide more harmonious teaching settings. Further development of data sources, more dimensional analysis indicators, and the development of knowledge about dynamic evolution patterns in educational public opinion can be among the aspects of future research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Chan, K. O. W., Ng, M. K. W., So, J. C. H., & Chan, V. C. W. (2021). Evaluation of Generic Competencies among Secondary School Leavers from the New Academic Structure for Senior Secondary Education in Hong Kong. *Public Administration and Policy, 24,* 182-194. https://doi.org/10.1108/pap-07-2020-0033

Dmoshinskaia, N., Gijlers, H., & de Jong, T. (2021). Learning from Reviewing Peers' Concept Maps in an Inquiry Context: Commenting or Grading, Which Is Better? *Studies in Educational Evaluation, 68,* Article 100959. https://doi.org/10.1016/j.stueduc.2020.100959

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent Advancements and Challenges of NLP-Based Sentiment Analysis: A State-of-the-Art Review. *Natural Language Processing Journal, 6,* Article 100059. https://doi.org/10.1016/j.nlp.2024.100059

Lei, C., & Tang, S. (2023). An Analysis of Hong Kong High School Curriculum with Implications for United Nations Sustainable Development Goals. *Smart Learning Environments, 10,* 1-18. https://doi.org/10.1186/s40561-023-00267-5

Lo, W. Y. W., & Li, D. (2023). Reimagining the Notion of Hong Kong as an Education Hub: National Imperative for Higher Education Policy. *International Journal of Educational Development, 103,* Article 102938. https://doi.org/10.1016/j.ijedudev.2023.102938

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment Analysis Methods, Applications, and Challenges: A Systematic Literature Review. *Journal of King Saud University-Computer and Information Sciences, 36,* Article 102048. https://doi.org/10.1016/j.jksuci.2024.102048

Tsui, K., Lee, C. J., Hui, K. S., Chun, W. D., & Chan, N. K. (2019). Academic and Career Aspiration and Destinations: A Hong Kong Perspective on Adolescent Transition. *Education Research International, 2019,* Article ID: 3421953. https://doi.org/10.1155/2019/3421953

Wang, Z., Huang, D., Cui, J., Zhang, X., Ho, S., & Cambria, E. (2025). A Review of Chinese Sentiment Analysis: Subjects, Methods, and Trends. *Artificial Intelligence Review, 58,* 1-35. https://doi.org/10.1007/s10462-024-10988-9

Yao, G. (2022). Deep Learning-Based Text Sentiment Analysis in Chinese International Promotion. *Security and Communication Networks, 2022,* Article ID: 7319656. https://doi.org/10.1155/2022/7319656

Yu, X., Wu, S., Chen, W., & Huang, M. (2021). Sentiment Analysis of Public Opinions on the Higher Education Expansion Policy in China. *Sage Open, 11,* 1-13. https://doi.org/10.1177/21582440211040778