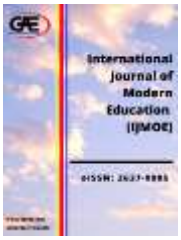
International Journal of Modern Education
EISSN: 2637-0905
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293



INTERNATIONAL JOURNAL OF MODERN EDUCATION (IJMOE)
www.gaexcellence.com/ijmoe

A DATA-CENTRIC STUDY OF ASPECT-BASED SENTIMENT ANALYSIS FOR SAFETY DISCOURSE IN TIKTOK TRAVEL CONTENT

Nur Shamilla Selamat[1]*, Jawahir Che Mustapha @ Yusuf[2], Juliana Jaafar[3]

[1]Department of Computer Science, Faculty of Engineering and Information Technology, Southern University College, Malaysia
nurshamilla@sc.edu.my https://orcid.org/0009-0009-9114-4598

[2]Information and Analytics Section, Malaysian Institute of Information Technology, Universiti Kuala Lumpur Malaysia
jawahir@unikl.edu.my https://orcid.org/0000-0003-4763-6797

[3]Artificial Intelligence Section, Malaysian Institute of Information Technology, Universiti Kuala Lumpur Malaysia
jjuliana@unikl.edu.my https://orcid.org/0000-0001-8761-8345

*Corresponding Author

Article Info:

Article history:
Received date: 08.02.2026
Revised date: 22.02.2026
Accepted date: 25.03.2026
Published date: 31.03.2026

To cite this document:

Selamat, N. S., Yusuf, J. C. M., & Jaafar, J. (2026). A Data-Centric Study of Aspect-Based Sentiment Analysis for Safety Discourse in TikTok Travel Content. *International Journal of Modern Education, 8*(29), 1277-1293.

Abstract:

Short-form social media platforms increasingly function as informal information systems through which safety perceptions are expressed and circulated. However, aspect-based sentiment analysis (ABSA) applied to safety discourse remains constrained by data sparsity, hierarchical label imbalance, and limited domain-specific resources. This study presents a data-centric investigation of ABSA for modelling safety-related discourse in TikTok comments associated with the #solotravel hashtag. A hierarchical annotation framework was developed comprising three primary safety aspects (physical/environmental, psychological/emotional, cultural/social), sixteen sub-aspects, and aspect-level sentiment polarity. Two datasets were constructed: a pilot dataset (402 comments) and an expanded dataset (2,362 comments), represented in multi-label exploded format. Classical baselines (Logistic Regression, SVM) and transformer architectures (DistilBERT, mBERT, XLM-R) were evaluated using five-fold cross-validation with macro-averaged metrics. Results demonstrate that annotation scale significantly influences model performance. At the pilot scale, classical models outperform transformer-based architectures. Following annotation expansion, transformer models surpass classical baselines for primary aspect classification, while classical models remain competitive for fine-grained subcategory and sentiment tasks. The results confirm that

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

dataset scale, label distribution, and hierarchical design exert stronger influence on ABSA effectiveness than architectural complexity alone. The findings provide empirical validation for data-centric system design principles and offer methodological guidance for annotation-driven analytical studies in emerging digital discourse domains.

DOI: 10.35631/IJMOE.829075

Keyword:

Aspect-Based Sentiment Analysis, Data-Centric AI, Safety Discourse, Social Media Mining, TikTok Analytics

Introduction

Short-form social media platforms such as TikTok and Instagram increasingly function as informal information systems through which travel-related experiences, emotions, and safety perceptions are exchanged in real time. Unlike traditional review platforms, these environments generate highly unstructured user-generated content (UGC) characterised by brevity, informality, emojis, and code-switching, which complicates systematic analysis.

TikTok has emerged as a highly influential platform shaping travel decision-making and destination perception, particularly among younger users, due to its algorithm-driven content exposure and high engagement rates (Fei et al., 2022; Hua et al., 2024; Huang et al., 2023; Zhang et al., 2023). However, the short-text, informal and context-dependent nature of TikTok comments introduces substantial linguistic variability and implicit meaning, posing challenges for computational analysis (Aziz et al., 2024).

To extract structured insights from such unstructured content, aspect-based sentiment analysis (ABSA) has been widely adopted within information systems and tourism research. ABSA enables fine-grained opinion mining by associating sentiment polarity with specific aspects rather than entire texts, thereby producing more actionable insights for decision-making (Chen et al., 2025; Hua et al., 2024; Jang et al., 2021; Zhu et al., 2022). However, prior studies consistently report that ABSA performance in short-form social media contexts is constrained primarily by data-related factors, including limited annotation scale, label sparsity, and class imbalance, rather than model architecture (Chen et al., 2025; Fei et al., 2022; Liu et al., 2025; Zhang et al., 2023). These challenges are particularly pronounced in safety-related travel discourse, where multiple dimensions of risk and emotion are often expressed implicitly within a single comment.

Despite substantial advances in sentiment analysis and ABSA, existing approaches remain insufficient for modelling safety discourse in TikTok travel content. First, most ABSA models are developed using structured datasets such as product reviews, limiting their effectiveness when applied to short, informal, and highly contextualised social media text (Ahmad et al.,

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

2025; Wankhade et al., 2024). Second, safety discourse in travel contexts is inherently multidimensional, involving emotional, cultural, and situational factors that are not adequately captured by conventional flat-aspect representations (Kandhro et al., 2024; Yin, 2024). Third, model performance is highly sensitive to dataset characteristics such as annotation quality, class imbalance, and dataset scale, yet these factors remain underexplored in ABSA research (Ali et al., 2022; Chauhan et al., 2024). These limitations motivate a data-centric investigation into ABSA for emerging social media discourse domains.

In response, this study adopts a data-centric perspective to examine how dataset scale, label structure, and annotation design influence the effectiveness of ABSA models for safety discourse analysis within TikTok travel content. Specifically, the study aims to (1) develop a hierarchical annotation framework for safety-related discourse, (2) evaluate the impact of annotation scale on model performance, and (3) compare classical and transformer-based approaches under varying data conditions. Using publicly available TikTok comments, the study constructs both a pilot dataset and an expanded annotated dataset of 2,362 comments, enabling controlled evaluation across three tasks: primary aspect classification, subcategory classification, and aspect-level sentiment prediction.

This study makes three key contributions to information systems and computational analytics research. First, it operationalises socio-cultural safety theory into a hierarchical, multi-label annotation framework tailored to short-form multilingual discourse. Second, it provides controlled empirical evidence isolating the causal impact of annotation scale on ABSA performance across modelling paradigms. Third, it offers a systematic comparison of classical and transformer-based architectures under identical data conditions, clarifying the conditions under which contextual modelling yields performance gains. Collectively, these contributions advance the design of data-centric analytical systems for safety-aware decision-making in short-form, multilingual digital environments.

The remainder of this paper is organised as follows. Section 2 reviews related work on ABSA, tourism analytics, and data-centric model design. Section 3 describes the data collection process, annotation framework, and experimental methodology. Section 4 presents and analyses the empirical results. Section 5 discusses implications for ABSA-driven information systems and safety analytics. Section 6 concludes the paper and outlines future research directions.

Literature Review

To situate this study within existing scholarship, this section reviews prior work on ABSA in tourism and social media contexts, with particular emphasis on data-centric challenges related to annotation design, dataset scale, and model performance.

Aspect-Based Sentiment Analysis in Tourism and Social Media

ABSA has been widely applied in tourism research to extract fine-grained opinions from UGC, particularly online reviews and social media posts. Prior studies primarily focus on service quality, accommodation, pricing, atmosphere, and destination attributes, linking aspect-level sentiment to tourist satisfaction, behavioural intention, or recommendation outcomes (Huang et al., 2023; Li et al., 2023; Mehra, 2023). Compared to document-level sentiment analysis, ABSA has been shown to produce more actionable insights for decision support in experiential

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

domains such as tourism (Chen et al., 2025; Hua et al., 2024; Jang et al., 2021; Zhu et al., 2022).

However, most tourism-oriented ABSA studies rely on long-form, review-based datasets where aspect boundaries and sentiment expressions are relatively explicit. The transfer of these methods to short-form social media environments remains underexplored (M. Abdelgwad et al., 2022; Perwira et al., 2025), despite growing evidence that platforms such as TikTok and Instagram play an increasingly influential role in shaping travel perceptions and decisions (Fei et al., 2022; Huang et al., 2023). Social media comments differ substantially from reviews in linguistic structure, exhibiting brevity, informality, emojis, and code-switching, which complicates aspect identification and sentiment attribution.

Data Scale, Annotation Design, and Class Imbalance in ABSA

A recurring challenge across ABSA robustness studies is that model performance is strongly influenced by dataset characteristics rather than architectural complexity alone. Annotation scale, label sparsity, and class imbalance are repeatedly identified as dominant factors limiting generalisation, particularly for fine-grained aspect classification (Chen et al., 2025; Fei et al., 2022; Liu et al., 2025; Zhang et al., 2023; Zhu et al., 2022). Rare or underrepresented aspects tend to be systematically misclassified, causing models to collapse predictions into dominant classes and reducing analytical usefulness.

To address this, recent ABSA datasets have shifted towards richer annotation schemes that allow multiple aspects and mixed sentiment polarity within a single instance. The MAMS dataset was explicitly introduced to overcome the limitations of earlier single-aspect corpora, enabling more realistic modelling of natural language opinion expression (Fei et al., 2022; Liu et al., 2025). In tourism-related applications, similar trends are observed, where improvements in performance are more strongly associated with annotation expansion and dataset rebalancing than with changes in model architecture (Huang et al., 2023; Perwira et al., 2025; Xu et al., 2024).

Classical and Transformer-Based Approaches

Classical machine learning methods, including Logistic Regression and Support Vector Machines with bag-of-words or TF–IDF representations, remain widely used as strong baselines in ABSA research (M. Abdelgwad et al., 2022; Perwira et al., 2025; Siddiqua et al., 2024). Comparative studies consistently report that these models perform competitively, and in some cases outperform transformer-based models, in low-resource or highly imbalanced settings (Chen et al., 2025; Zhang et al., 2023).

Transformer-based models such as BERT and its multilingual variants have demonstrated strong performance in ABSA tasks, especially in under-resourced languages and code-switched environments (M. Abdelgwad et al., 2022; Perwira et al., 2025). Nevertheless, empirical evidence suggests that such models require sufficiently large and balanced datasets to reliably surpass classical baselines. When annotation is sparse or skewed, transformers may exhibit instability and reduced macro-averaged performance, especially for fine-grained aspect categorisation (Hua et al., 2024; Zhu et al., 2022).

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

Sentiment vs. Aspect Learnability

Across ABSA literature, sentiment polarity classification is consistently reported as more learnable than aspect identification, especially in short texts (Fei et al., 2022; Zhang et al., 2023; Zhu et al., 2022). Sentiment cues tend to be linguistically explicit, whereas aspect boundaries, especially in domains such as safety, are often implicit, contextual, and multi-dimensional. This observation is replicated in tourism and social media studies, where models achieve higher and more stable performance on sentiment prediction than on nuanced aspect or sub-aspect detection (Huang et al., 2023; Mehra, 2023).

Collectively, existing studies demonstrate that ABSA performance is primarily determined by dataset characteristics rather than model complexity. However, prior research has largely focused on general domains or long-form textual data, with limited attention to safety-oriented discourse in short-form, multilingual social media environments. In particular, there is a lack of controlled, scale-sensitive comparisons across modelling approaches that explicitly examine how annotation density and label structure influence performance. This gap limits understanding of when advanced contextual models meaningfully outperform classical baselines under realistic data constraints.

Recent research increasingly emphasises data-centric approaches in natural language processing, demonstrating that improvements in dataset quality, annotation consistency, and label design can substantially enhance model performance (Chauhan et al., 2024; Drašković & Milanović, 2025). At the same time, studies on social media analytics highlight the challenges of analysing short-form, informal, and context-dependent text, particularly in platforms such as TikTok, where discourse is highly dynamic and often multimodal (Aziz et al., 2024; Chu et al., 2022). Despite these developments, ABSA research remains predominantly model-centric and focused on traditional domains, with limited investigation into how data characteristics influence performance in emerging discourse environments. This study addresses this gap through a data-centric experimental design applied to safety-related travel discourse.

Methodology

Research Design

This study adopts a data-centric experimental design to examine how dataset scale, annotation structure, and label distribution influence the performance of ABSA models in short-form social media contexts. Recent studies in natural language processing increasingly demonstrate that data quality and annotation design exert a stronger influence on model effectiveness than architectural complexity, particularly in fine-grained tasks such as ABSA (Tao & Fang, 2020; Zhang et al., 2023). In response, this study isolates the effect of annotation scale by constructing two datasets, specifically a pilot dataset (402 comments) and an expanded dataset (2,362 comments), while maintaining identical preprocessing, modelling configurations, and evaluation protocols across both datasets.

By controlling for algorithmic and procedural variation, the design enables attribution of performance differences to dataset characteristics such as annotation density, class distribution, and hierarchical label coverage. This approach further enables examination of architectural crossover effects, specifically whether transformer-based models outperform classical baselines only after sufficient annotation scale is achieved, thereby situating the study within

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

emerging data-centric AI paradigms. The overall data-centric ABSA pipeline, integrating data collection, annotation, and multi-architecture evaluation, is illustrated in Figure 1.

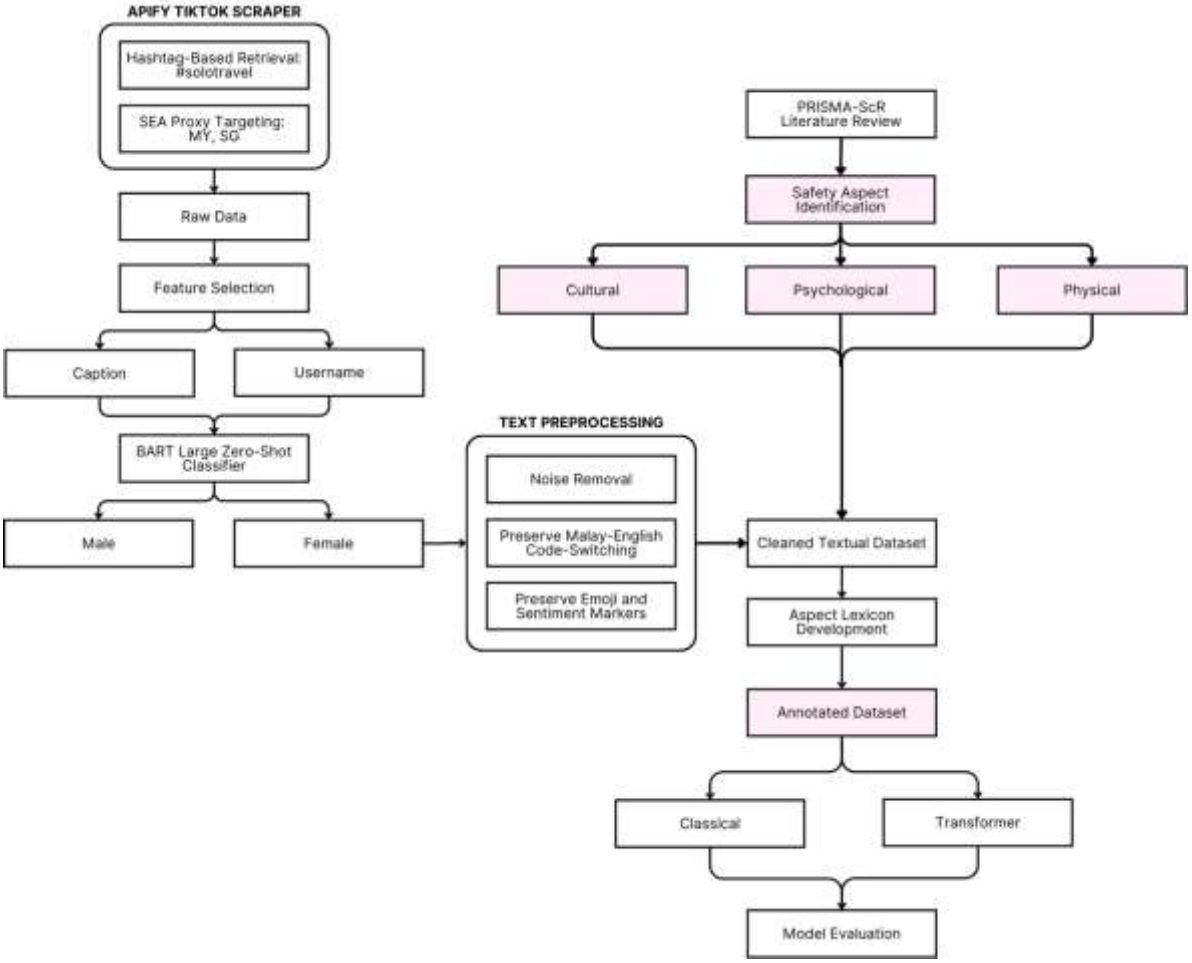


Figure 1. Data-Centric ABSA Pipeline for Safety-Related Travel Discourse

Conceptual Basis and Annotation Framework

The annotation framework is grounded in a systematic literature review conducted in accordance with PRISMA-ScR guidelines, which synthesised research on gendered safety in solo female travel (Selamat et al., 2026). Prior studies consistently conceptualise safety as a multi-dimensional construct encompassing physical and environmental risks, psychological and emotional responses, and cultural and social constraints (Chowdhury et al., 2023; Ghadban et al., 2023; Ison et al., 2023). These dimensions were operationalised into three primary aspect domains, which are physical and environmental safety, psychological and emotional safety, and cultural and social constraints, alongside a "none" category for non-safety-related discourse.

Each domain was further refined into inductively derived sub-aspects reflecting recurrent discourse patterns, enabling fine-grained and interpretable classification. In addition, each instance was annotated with aspect-level sentiment polarity (positive, negative, neutral, or mixed) and categorised as either explicit or implicit safety discourse. This hierarchical and multi-dimensional representation aligns with recent ABSA research advocating domain-specific taxonomies and richer annotation schemes to improve analytical precision and ecological validity (Alotaibi & Nadeem, 2025; Dai et al., 2025). The structured taxonomy

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

enables systematic comparison across modelling levels, allowing evaluation of how annotation scale differentially affects coarse-grained and fine-grained aspect classification under both lexical and contextual architectures.

Annotation Procedure and Reliability

Annotation was guided by a structured codebook derived from the literature review, specifying operational definitions, inclusion and exclusion criteria, and decision rules for handling implicit expressions, multilingual content, and informal linguistic features such as emojis and code-switching. Annotators underwent a structured training process involving codebook familiarisation, independent pilot annotation, and iterative calibration through discussion of disagreements. This process ensured conceptual alignment and reduced ambiguity, particularly for complex cases involving overlapping sub-aspects or implicit safety cues, which are known challenges in short-form discourse analysis (Fei et al., 2022; Hua et al., 2024).

Inter-annotator agreement was evaluated on 100 randomly sampled comments using Cohen's Kappa coefficient. The results indicate substantial to near-perfect agreement, with $\kappa = 0.84$ for primary aspect classification, $\kappa = 0.76$ for sub-aspect classification, and $\kappa = 0.88$ for sentiment polarity. In practical terms, only one to two instances per 100 comments required relabelling after consensus discussion, indicating high annotation consistency. Disagreements were resolved through adjudication, and the refined guidelines were applied to the full dataset, ensuring reliability and internal validity.

Data Collection and Preprocessing

UGC was collected from TikTok using the Apify TikTok Scraper, targeting publicly available posts associated with the hashtag #solotravel. Geo-targeted proxy configurations were applied to prioritise Southeast Asian content, and complete comment threads were retrieved to preserve conversational context. To enhance relevance to solo female travel discourse, creator gender was probabilistically inferred using a zero-shot classification model applied to publicly available self-descriptions, following an established approach by Pitanatri et al. (2025). This inference was used solely for sampling relevance and not for identity profiling or demographic analysis. Only content likely generated by individual travellers, rather than commercial accounts, was retained.

Preprocessing involved schema standardisation, removal of URLs and user mentions, duplicate filtering, and whitespace normalisation, while preserving emojis, informal expressions, and code-switching due to their semantic importance in sentiment expression (Dai et al., 2025; Hua et al., 2024). The annotation scheme supports multi-label assignment, allowing multiple aspect–sentiment pairs within a single comment. The dataset was subsequently transformed into an exploded format, where each row represents a single aspect–sentiment pair, consistent with recent ABSA modelling practices (Tao & Fang, 2020; Zhang et al., 2023).

Experimental Setup

The experimental framework evaluates both classical machine learning models and transformer-based architectures under identical conditions. Classical baselines include Logistic Regression with bag-of-words features and Support Vector Machines with TF–IDF representations, which have been shown to perform robustly in low-resource and imbalanced

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

ABSA settings (Motevalli et al., 2025; Siddiqua et al., 2024). Transformer-based models include DistilBERT, multilingual BERT (mBERT), and XLM-RoBERTa (XLM-R), enabling evaluation of contextual representations in multilingual and code-switched environments.

Transformer models were fine-tuned using a sequence classification framework with cross-entropy loss, trained for three epochs using AdamW optimisation with a learning rate of 2e-5 and a batch size of 16. Early stopping based on validation loss was applied to prevent overfitting. All models were evaluated across three tasks: primary aspect classification, sub-aspect classification, and aspect-level sentiment classification, ensuring consistency across architectures.

Evaluation Strategy

Model performance was assessed using five-fold stratified cross-validation to preserve label distribution across folds, a standard approach for handling imbalanced classification tasks (Ramasamy & Elangovan, 2024; Zhang et al., 2023). Evaluation metrics include accuracy, precision, recall, and macro-averaged F1-score, with macro-F1 prioritised due to its sensitivity to minority classes and suitability for hierarchical multi-label classification.

To isolate the impact of annotation scale, performance on the pilot dataset was directly compared with performance on the expanded dataset while maintaining identical experimental conditions. This scale-sensitive evaluation enables attribution of performance differences to dataset characteristics rather than model variation and supports analysis of when contextual models begin to outperform lexical baselines. Such an approach aligns with recent findings that performance improvements in ABSA are often driven by dataset expansion and annotation quality rather than architectural substitution alone (Tao & Fang, 2020; Yang & Liu, 2025).

Results

This section reports and interprets the performance of classical and transformer models for ABSA across two annotation scales, which include an initial pilot dataset of 402 annotated comments and an expanded annotated dataset of 2,362 comments. The comparative analysis is designed to examine how annotation scale, class imbalance, and hierarchical multi-label representation influence the learnability of safety-related aspects and sentiments in short-form social media discourse. In line with recent data-centric perspectives, the results are interpreted not only in terms of model performance but also with respect to underlying dataset characteristics that shape model behaviour (Tao & Fang, 2020; Zhang et al., 2023).

Results on the Pilot Dataset

Table 1 summarises the performance of classical and transformer models across three tasks: primary aspect classification, subcategory aspect classification, and aspect-level sentiment classification. At this annotation scale, performance is constrained by label sparsity and severe class imbalance, which are known to limit generalisation in fine-grained ABSA tasks (Fei et al., 2022; Liu et al., 2025).

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

Table 1. Performance of Classical and Transformer Models on the Pilot Dataset

Task	Model	Macro Precision	Macro Recall	Macro F1	Accuracy
Primary Aspect	Logistic Regression (BoW)	0.414	0.383	0.378	0.737
	SVM (TF-IDF)	0.405	0.401	**0.388**	0.734
	DistilBERT	0.332	0.380	0.352	0.724
	mBERT	0.329	0.398	0.354	0.711
	XLM-R	0.191	0.267	0.221	0.659
Subcategory Aspect	Logistic Regression (BoW)	0.171	0.167	0.163	0.712
	SVM (TF-IDF)	0.249	0.221	**0.220**	0.712
	DistilBERT	0.100	0.114	0.101	0.664
	mBERT	0.123	0.157	0.134	0.689
	XLM-R	0.065	0.100	0.079	0.649
Aspect-Level Sentiment	Logistic Regression (BoW)	0.564	0.503	0.519	0.717
	SVM (TF-IDF)	0.686	0.510	**0.545**	0.719
	DistilBERT	0.286	0.267	0.218	0.612
	mBERT	0.304	0.329	0.307	0.634
	XLM-R	0.166	0.252	0.191	0.595

For primary aspect classification, classical linear models demonstrate superior performance. SVM achieves the highest macro-F1 score (0.388), followed closely by Logistic Regression (0.378), while transformer models exhibit weaker and less stable performance, particularly for minority classes. Although accuracy values exceed 0.70, this is largely driven by the dominance of the "none" class, which constitutes approximately 65% of the dataset. This discrepancy highlights the limitations of accuracy as an evaluation metric in imbalanced multi-class settings and reinforces the use of macro-F1 as a more reliable indicator of discriminative capability (Ramasamy & Elangovan, 2024; Zhang et al., 2023). The relatively stronger performance of classical models at this scale is consistent with prior findings that lexical models generalise more effectively under data scarcity due to their reliance on surface-level features rather than contextual representations (Hua et al., 2024; Siddiqua et al., 2024).

Subcategory aspect classification yields the weakest performance across all models, with macro-F1 scores remaining below 0.22. This reflects extreme sparsity across the sixteen fine-grained sub-aspects, where many categories have insufficient training instances to support reliable learning. While SVM performs comparatively better (0.220), both precision and recall remain low, indicating limited class separability. Transformer architectures perform substantially worse, indicating that contextual models are particularly sensitive to insufficient data in hierarchical classification settings. This observation aligns with existing literature indicating that fine-grained aspect detection is highly dependent on annotation density and balanced label distribution (Chen et al., 2025; Liu et al., 2025).

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

Aspect-level sentiment classification produces comparatively stronger results, with SVM achieving a macro-F1 score of 0.545. This performance gap between sentiment and aspect classification is consistent with established findings that sentiment polarity is generally more lexically explicit and therefore easier to learn, particularly in short-form text (Zhang et al., 2023; Zhu et al., 2022). In the context of TikTok comments, sentiment is often conveyed through explicit lexical markers or emojis, reducing reliance on complex contextual modelling. Overall, the pilot dataset results indicate that limited annotation density constrains model performance, particularly for hierarchical and fine-grained aspect detection tasks.

Results on the Expanded Dataset

Table 2 presents results after expanding the annotated dataset to 2,362 comments using a multi-label representation. Across all tasks and architectures, substantial improvements are observed, confirming that annotation scale and label coverage are primary determinants of ABSA performance.

Table 2. Performance of Classical and Transformer Models on the Expanded Dataset

Task	Model	Macro Precision	Macro Recall	Macro F1	Accuracy
Primary Aspect	Logistic Regression (BoW)	0.688	0.628	0.650	0.757
	SVM (TF-IDF)	0.682	0.635	0.653	0.753
	DistilBERT	0.732	0.660	0.678	0.793
	mBERT	0.724	0.687	**0.698**	0.784
	XLM-R	0.725	0.642	0.643	0.779
Subcategory Aspect	Logistic Regression (BoW)	0.433	0.347	0.370	0.664
	SVM (TF-IDF)	0.427	0.404	**0.408**	0.665
	DistilBERT	0.307	0.274	0.277	0.685
	mBERT	0.350	0.317	0.311	0.680
	XLM-R	0.207	0.199	0.184	0.626
Aspect-Level Sentiment	Logistic Regression (BoW)	0.717	0.698	0.698	0.751
	SVM (TF-IDF)	0.734	0.685	**0.699**	0.781
	DistilBERT	0.620	0.555	0.566	0.735
	mBERT	0.616	0.552	0.559	0.730
	XLM-R	0.635	0.532	0.532	0.729

For primary aspect classification, transformer models now outperform classical baselines, indicating a clear scale-dependent shift in model effectiveness. mBERT achieves the highest macro-F1 score (0.698), followed by DistilBERT (0.678), while classical models also improve significantly (SVM: 0.653; Logistic Regression: 0.650). The improved balance between precision and recall across models suggests enhanced minority-class detection, reflecting increased representation of previously undersampled categories. This crossover effect provides empirical support for prior studies demonstrating that transformer-based models require

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

sufficient training data to realise their contextual representation advantages (Hua et al., 2024; Zhu et al., 2022). It further reinforces the argument that model superiority in ABSA is conditional upon dataset scale rather than inherent to architecture (Chen et al., 2025; Yang & Liu, 2025).

Subcategory aspect classification also improves, although performance remains moderate relative to other tasks. SVM achieves the highest macro-F1 score (0.408), followed by Logistic Regression (0.370), while transformer models improve but do not surpass classical baselines. This suggests that even with increased data, fine-grained sub-aspects remain sensitive to class imbalance and semantic overlap. The persistence of this pattern indicates that lexical separability may still dominate at the subcategory level, particularly when contextual distinctions are subtle or inconsistently expressed. Similar findings have been reported in multi-aspect datasets such as MAMS, where rare categories remain difficult to model despite increased annotation (Fei et al., 2022).

Aspect-level sentiment classification achieves the highest overall performance across all tasks, with SVM reaching a macro-F1 score of 0.699 and accuracy of 0.781. Transformer models also improve but remain slightly inferior to classical baselines. This confirms that sentiment polarity is consistently more learnable than hierarchical aspect structures, particularly in short-form discourse where emotional cues are explicitly encoded (Chen et al., 2025; Mehra, 2023). The relatively smaller performance gap between models suggests that sentiment classification benefits less from contextual modelling compared to aspect detection.

Discussion

Figure 2 synthesises macro-F1 performance across models and annotation scales, revealing three consistent patterns. First, annotation expansion yields substantial performance improvements across all architectures, confirming that dataset scale is a primary driver of model effectiveness. Second, the magnitude of improvement varies by task, with primary aspects benefiting most from increased data, while subcategory classification remains constrained by structural complexity and imbalance. Third, a scale-dependent crossover effect is observed, where transformer models surpass classical baselines only after sufficient annotation density is achieved.

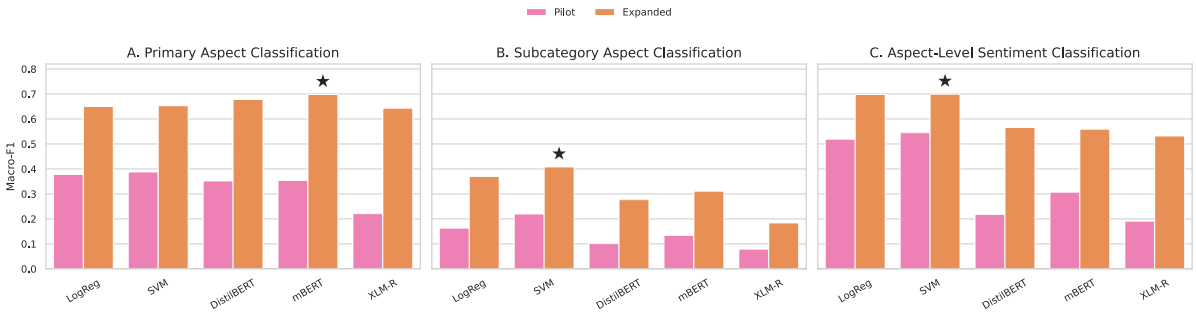


Figure 2. Macro-F1 Performance Comparison Across Classical and Transformer Models for Pilot and Expanded Annotated Datasets

The observed crossover effect has important methodological implications. Under low-resource conditions, classical models offer greater stability and efficiency due to their reliance on surface-level lexical features. However, as the dataset scale increases, transformer models

become more effective in capturing contextual and implicit relationships, particularly in multilingual and code-switched environments (Perwira et al., 2025). This finding aligns with prior robustness studies suggesting that contextual models require a critical threshold of annotation density to outperform simpler baselines (Fei et al., 2022; Zhang et al., 2023).

The persistent difficulty in subcategory classification further highlights the limitations of hierarchical ABSA in complex discourse domains. Despite dataset expansion, performance remains moderate, suggesting that fine-grained categories are inherently more difficult to model due to semantic overlap and implicit expression. This supports prior work emphasising the importance of taxonomy design and label consolidation in improving model performance (Dai et al., 2025; Hua et al., 2024).

Overall, the findings provide strong empirical support for a data-centric interpretation of ABSA system behaviour. Performance improvements are primarily driven by annotation expansion, improved label distribution, and hierarchical clarity rather than architectural innovation. Classical models remain competitive under low-resource conditions, while transformer models demonstrate clear advantages only when sufficient annotated data is available. These results align with broader calls in ABSA research to prioritise dataset engineering and annotation design as first-order determinants of system performance (Chauhan et al., 2024; Tao & Fang, 2020).

Beyond performance evaluation, this study contributes a domain-specific, hierarchical annotation framework for modelling safety discourse in short-form social media. By operationalising physical, psychological, and cultural safety dimensions into structured computational categories, the framework bridges ABSA methodologies with socio-cultural research and supports the development of scalable, data-driven safety analytics systems for emerging digital platforms.

Despite these findings, the study is limited by residual class imbalance and the moderate size of the annotated dataset, particularly at the subcategory level, which may constrain the generalisability of the results to broader or more diverse social media contexts. In addition, the reliance on text-only analysis without incorporating multimodal signals such as video or audio content may limit the ability to fully capture the richness of safety discourse in TikTok environments.

Conclusion

This study provides a data-centric examination of how annotation scale and hierarchical multi-label design influence ABSA performance in modelling safety-related discourse within short-form social media. Using TikTok comments associated with solo female travel, the study systematically compared classical and transformer-based models across two annotation scales while holding modelling configurations constant. This controlled design enabled isolation of the effects of dataset size, label distribution, and hierarchical representation on model performance.

The findings demonstrate that ABSA performance is fundamentally governed by data quality and dataset scale rather than architectural complexity alone. Substantial and consistent improvements were observed across all tasks following annotation expansion, confirming that increased annotation density enhances class representation, reduces sparsity, and stabilises

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

model learning. A clear scale-dependent crossover effect was identified, whereby classical models outperform under low-resource conditions, while transformer architectures such as mBERT achieve superior performance only after sufficient annotation coverage is reached. This provides strong empirical support for data-centric AI principles, reinforcing dataset engineering as the primary driver of analytical reliability in ABSA.

The study further contributes a structured, safety-oriented hierarchical annotation framework grounded in socio-cultural theory, enabling the systematic operationalisation of multi-dimensional safety discourse into computationally tractable categories. In addition, the results clarify task-level learnability, showing that sentiment polarity is consistently easier to model than hierarchical aspect structures, while fine-grained sub-aspect classification remains highly sensitive to imbalance and label sparsity.

From a practical standpoint, these findings inform the design of analytical information systems for safety monitoring and decision support in tourism contexts, particularly within informal, multilingual, and short-text environments. The results suggest that investment in annotation quality, dataset expansion, and taxonomy design yields greater performance gains than incremental model optimisation.

Despite these contributions, limitations remain, including residual class imbalance, platform-specific data collection, and the exclusion of multimodal signals inherent to TikTok content. Future research should prioritise large-scale, balanced annotation, cross-platform validation, and multimodal integration to further improve robustness. Overall, this study reinforces that reliable ABSA in emerging discourse domains depends primarily on the quality, structure, and scale of data, providing both methodological clarity and practical guidance for future ABSA research and system development.

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

Acknowledgements: The authors would like to express their sincere gratitude to Southern University College and Malaysian Institute of Information Technology, Universiti Kuala Lumpur for providing the necessary resources and support throughout the course of this research. Special appreciation is extended to supervisors who contributed valuable insights and constructive feedback, which greatly enhanced the quality of this paper.

Funding Statement: No Funding

Conflict of Interest Statement: The authors declare that there is no conflict of interest regarding the publication of this paper. All authors have contributed to this work and approved the final version of the manuscript for submission to the International Journal of Modern Education (IJMOE).

Ethics Statement: This study did not involve direct interaction with any human participants, animals, or sensitive data requiring ethical approval. The dataset consisted solely of publicly available TikTok comments collected in accordance with platform accessibility guidelines. No private, restricted, or sensitive personal data were accessed. Usernames and identifiable information were removed during preprocessing to ensure anonymity. The authors confirm that the research was conducted in accordance with accepted academic integrity and ethical publishing standards.

Author Contribution Statement: All authors contributed significantly to the development of this manuscript. Dr Jawahir and Ts Dr Juliana were responsible for the conceptualization, critical revision of the manuscript, and overall supervision of the study. Nur Shamilla contributed to literature review, data collection, analysis, and interpretation of results. All authors read and approved the final version of the manuscript prior to submission.

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

References

Ahmad, W., Khan, H. U., Alarfaj, F., & Alreshoodi, M. (2025). Aspect-Based Sentiment Analysis: A Comprehensive Review and Open Research Challenges. *IEEE Access*, *13*, 65138–65182. https://doi.org/10.1109/access.2025.3555744

Ali, T., Omar, B., & Soulaimane, K. (2022). Analyzing tourism reviews using an LDA topic-based sentiment analysis approach. *MethodsX*, *9*, 101894. https://doi.org/10.1016/j.mex.2022.101894

Alotaibi, A., & Nadeem, F. (2025). An Unsupervised Integrated Framework for Arabic Aspect-Based Sentiment Analysis and Abstractive Text Summarization of Traffic Services Using Transformer Models. *Smart Cities*, *8*(2). https://doi.org/10.3390/smartcities8020062

Aziz, K., Ji, D., Chakrabarti, P., Chakrabarti, T., Iqbal, M. S., & Abbasi, R. (2024). Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection. *Scientific Reports*, *14*, 14646. https://doi.org/10.1038/s41598-024-61886-7

Chauhan, A., Sharma, A., & Mohana, R. (2024). A Pre-Trained Model for Aspect-based Sentiment Analysis Task: Using Online Social Networking. In *Procedia Computer Science* (Vol. 233, pp. 35–44). Elsevier B.V. https://doi.org/10.1016/j.procs.2024.03.193

Chen, X., Xie, H., Tao, X., Wang, F. L., Zhang, D., & Dai, H.-N. (2025). A Computational Analysis of Aspect-based Sentiment Analysis Research Through Bibliometric Mapping and Topic Modeling. *Journal of Big Data*, *12*(1), 40. https://doi.org/10.1186/s40537-025-01068-y

Chowdhury, S., Patel, P., Giridharan, V., & Ceccato, V. (2023). Formation of Fear and Adaptive Behavior in Young Ethnic Minority Women Riding Public Transport. *Transportation Research Record*. https://doi.org/10.1177/03611981231182712

Chu, M., Chen, Y., Yang, L., & Wang, J. (2022). Language interpretation in travel guidance platform: Text mining and sentiment analysis of TripAdvisor reviews. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1029945

Dai, W., Kong, W., Shang, T., Feng, J., Wu, J., & Qu, T. (2025). Guideline for Novel Fine-Grained Sentiment Annotation and Data Curation: A Case Study. *Expert Systems*. https://doi.org/10.1111/exsy.70022

Drašković, D., & Milanović, S. (2025). Aspect-based sentiment analysis of user-generated content from a microblogging platform. *Journal of Big Data*, *12*(1), 186. https://doi.org/10.1186/s40537-025-01244-0

Fei, H., Chua, T.-S., Li, C., Ji, D., Zhang, M., & Ren, Y. (2022). On the Robustness of Aspect-based Sentiment Analysis: Rethinking Model, Data, and Training. *ACM Trans. Inf. Syst.*, *41*(2), 50:1-50:32. https://doi.org/10.1145/3564281

Ghadban, S., Kamar, R., & Haidar, R. (2023). Decoding International Solo Women Travelers' Experience: A Qualitative Analysis of User-generated Videos. *Journal of Outdoor Recreation and Tourism*, *44*, 100648. https://doi.org/10.1016/j.jort.2023.100648

Hua, Y. C., Denny, P., Wicker, J., & Taskova, K. (2024). A Systematic Review of Aspect-based Sentiment Analysis: Domains, Methods, and Trends. *Artificial Intelligence Review*, *57*(11), 296. https://doi.org/10.1007/s10462-024-10906-z

Huang, S., Wu, X., Wu, X., & Wang, K. (2023). Sentiment Analysis Algorithm Using Contrastive Learning and Adversarial Training for POI Recommendation. *Social Network Analysis and Mining*. https://link.springer.com/article/10.1007/s13278-023-01076-x

Ison, J., Forsdike, K., Henry, N., Hooker, L., & Taft, A. (2023). "You're just constantly on alert": Women and Gender-Diverse People's Experiences of Sexual Violence on Public Transport. *Journal of Interpersonal Violence*, *38*(21–22), 11617–11641. https://doi.org/10.1177/08862605231186123

Jang, H., Rempel, E., Roth, D., Carenini, G., & Janjua, N. Z. (2021). Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis. *Journal of Medical Internet Research*, *23*(2), e25431. https://doi.org/10.2196/25431

Kandhro, I. A., Ali, F., Uddin, M., Kehar, A., & Manickam, S. (2024). Exploring aspect-based sentiment analysis: An in-depth review of current methods and prospects for advancement. In *Knowledge and Information Systems* (Vol. 66, Issue 7, pp. 3639–3669). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s10115-024-02104-8

Li, H., Yu, B. X. B., Li, G., & Gao, H. (2023). Restaurant Survival Prediction Using Customer-generated Content: An Aspect-based Sentiment Analysis of Online Reviews. *Tourism Management*, *96*, 104707. https://doi.org/10.1016/j.tourman.2022.104707

Liu, J., Tian, Y., & Song, Y. (2025). *Balanced Training Data Augmentation for Aspect-Based Sentiment Analysis* (arXiv:2507.09485). arXiv. https://doi.org/10.48550/arXiv.2507.09485

M. Abdelgwad, M., A Soliman, T. H., I.Taloba, A., & Farghaly, M. F. (2022). Arabic Aspect-based Sentiment Analysis Using Bidirectional GRU-based Models. *Journal of King Saud University - Computer and Information Sciences*, *34*(9), 6652–6662. https://doi.org/10.1016/j.jksuci.2021.08.030

Mehra, P. (2023). Unexpected Surprise: Emotion Analysis and Aspect-based Sentiment Analysis (ABSA) of User-generated Comments to Study Behavioral Intentions of Tourists. *Tourism Management Perspectives*, *45*, 101063. https://doi.org/10.1016/j.tmp.2022.101063

Motevalli, M. M., Sohrabi, M. K., & Yaghmaee, F. (2025). Aspect-based Sentiment Analysis: A Dual-task Learning Architecture Using Imbalanced Maximized-area Under the Curve Proximate Support Vector Machine and Reinforcement Learning. *Information Sciences*, *689*, 121449. https://doi.org/10.1016/j.ins.2024.121449

Perwira, R. I., Purnamasari, D. I., Permadi, V. A., & Agusdin, R. P. (2025). Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel User-Generated Content. *Journal of Information Systems Engineering and Business Intelligence*, *11*(1), 30–40. https://doi.org/10.20473/jisebi.11.1.30-40

Pitanatri, P. D. S., Adnyani, N. W. G., Kartini, L. P., & Valeri, M. (2025). Travel Motivations, Preferences, and Characteristics of Women Solo Travelers in Bali. *Journal of Applied Sciences in Travel and Hospitality*, *8*(1), Article 1. https://doi.org/10.31940/jasth.v8i1.63-78

Ramasamy, M., & Elangovan, M. (2024). Optimized Neural Attention Mechanism for Aspect-based Sentiment Analysis Framework with Optimal Polarity-based Weighted Features. *Knowledge and Information Systems*, *66*, 2501–2535.

Selamat, N. S., Che Mustapha, J., & Jaafar, J. (2026). Mapping Gendered Safety in Solo Female Travel: A Scoping Review of Computational Approaches and Digital Discourse. *International Journal for Studies on Children, Women, Elderly and Disabled*, *26*.

Siddiqua, A., Bindumathi, V., Raghu, G., & Vamsi Bhargav, Y. S. (2024). Aspect-based Sentiment Analysis (ABSA) using Machine Learning Algorithms. *2024 Third International Conference on Distributed Computing and Electrical Circuits and*

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume: 8 Issue: 29 (March 2026) PP. 1277-1293

Electronics (ICDCECE), 1–6. https://doi.org/10.1109/ICDCECE60827.2024.10549140

Tao, J., & Fang, X. (2020). Toward Multi-label Sentiment Analysis: A Transfer Learning-based Approach. *Journal of Big Data*, *7*(1), 1. https://doi.org/10.1186/s40537-019-0278-0

Wankhade, M., Kulkarni, C., & Rao, A. C. S. (2024). A survey on aspect base sentiment analysis methods and challenges. *Applied Soft Computing*, *167*, 112249. https://doi.org/10.1016/j.asoc.2024.112249

Xu, C., Wang, M., Ren, Y., & Zhu, S. (2024). *Enhancing Aspect-based Sentiment Analysis in Tourism Using Large Language Models and Positional Information* (arXiv:2409.14997). arXiv. https://doi.org/10.48550/arXiv.2409.14997

Yang, Y., & Liu, F. (2025). Research on Aspect-based Sentiment Analysis of Homestay Online Comments Based on BERT-BiLSTM-Multi-Head Attention. *International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2024)*, *13560*, 264–269. https://doi.org/10.1117/12.3061617

Yin, S. (2024). The Current State and Challenges of Aspect-Based Sentiment Analysis. *Applied and Computational Engineering*, *114*, 25–31. https://doi.org/10.54254/2755-2721/2024.18197

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2023). A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *IEEE Transactions on Knowledge and Data Engineering*, *35*(11), 11019–11038. https://doi.org/10.1109/TKDE.2022.3230975

Zhu, L., Xu, M., Bao, Y., Xu, Y., & Kong, X. (2022). Deep Learning for Aspect-based Sentiment Analysis: A Review. *PeerJ Computer Science*. https://peerj.com/articles/cs-1044/