Online Information Extraction System (EAOS) for Social Media Comments

Khanh-Hung Huynh, Phuc Bui Hoang Gia, Huyen Dinh Doan My,
Phi-Long Nguyen, Thien-Vu Nguyen-Ho, Trong-Hop Do

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

Online Information Extraction System (EAOS) for Social Media Comments

Khanh-Hung Huynh[1,2], **Phuc Bui Hoang Gia**[1,2], **Huyen Dinh Doan My**[1,2],
Phi-Long Nguyen[1,2], **Thien-Vu Nguyen-Ho**[1,2], **Trong-Hop Do**[1,2],

[1]University of Information Technology,
[2]Vietnam National University Ho Chi Minh city
Corresponding author: hopdt@uit.edu.vn

Abstract

An EAOS (Entity–Aspect–Opinion–Sentiment) information extraction system is proposed to analyze Vietnamese user comments on the YouTube platform. This work is the first to address a unified multi-task framework for Vietnamese social media content, jointly extracting entities, aspect categories, opinion spans, and sentiment—contrasting with prior studies that focused only on single tasks such as sentiment analysis or aspect-based sentiment analysis. The data is collected from highly interactive videos containing various opinions about products, services, and social topics. The system includes an offline module for data preprocessing, model construction, and training using deep learning techniques, and a prediction module for applying the trained model to real-world data. In the offline phase, the pretrained PhoBERT model is used to encode the contextual semantics of text, followed by classification layers that simultaneously predict sentiment, aspect category, and the start–end positions of both entity and opinion spans within each sentence. The dataset is normalized and manually annotated in the EAOS format, then divided into training, development, and testing sets. Experimental results show that PhoBERT outperforms traditional baseline methods, especially in accurately identifying aspect-related and opinion-bearing expressions. To ensure practical applicability, the trained models are deployed on an Apache Spark–based streaming architecture, enabling the system to process large-scale and continuous social media data in real time. The system is designed to be flexible and scalable, enabling adaptation to other social media platforms such as Facebook or TikTok. These findings emphasize both the pioneering nature of this research in Vietnamese multi-task information extraction and the effectiveness of transformer-based models in extracting complex information from Vietnamese social media content.

1 Introduction

In the current digital age, social media serves as a significant medium for expressing public opinions, emotions, and evaluations regarding products, services, and public figures. YouTube, as one of the most popular platforms, generates a continuous stream of user comments, providing a rich and real-time data source for opinion mining and sentiment analysis. The task addressed in this study involves the extraction of EAOS (Entity – Aspect – Opinion – Sentiment) structures from Vietnamese YouTube comments, with the goal of identifying the mentioned subject, the specific aspect being discussed, the expressed opinion, and the corresponding sentiment. To the best of our knowledge, this is the first work to address a unified multi-task framework for Vietnamese social media content, jointly extracting entities, aspect categories, opinion spans, and sentiment, rather than focusing on a single task as in prior studies such as sentiment analysis or aspect-based sentiment analysis.

The EAOS framework includes four core components: Entity, which refers to the object being mentioned (such as an artist, program, or product); Aspect, which denotes the aspect or topic associated with the entity (such as appearance, personality, or expertise); Opinion, which is the textual phrase reflecting the user's judgment; and Sentiment, which captures the polarity of the opinion (positive, negative, or neutral). The task requires models to detect and correctly associate these elements within each comment, considering that a single sentence may contain multiple EAOS tuples or express components in implicit ways.

Unlike conventional sentiment analysis approaches that focus on classifying the sentiment of entire sentences, the EAOS task demands a more fine-grained understanding of sentence semantics and structure. Additionally, the real-time and large-scale nature of YouTube comments introduces the need for a system capable of processing data instantly as it is posted. To meet both the accuracy and scalability requirements, our proposed EAOS extraction system integrates deep learning techniques with Apache Spark Streaming, enabling large-scale, real-time processing of social media data. This architecture allows the trained models to detect entities, aspects, opinions, and sentiments in newly submitted comments without delay.

Designed for applications in social media monitoring and decision support, the system not only captures general public sentiment but also provides detailed insights into specific aspects of interest. By combining a pioneering multi-task approach, a manually annotated EAOS dataset for Vietnamese, and a scalable streaming-based deployment, this work lays the foundation for practical, real-time EAOS extraction from Vietnamese-language social media and contributes to advancing semantic understanding technologies in this domain.

2 Related Work

2.1 EAOS Extraction Systems

Over the past decade, the field of Aspect-Based Sentiment Analysis (ABSA) has attracted significant attention from the research community, with various approaches proposed to jointly extract and associate semantic components such as entities, aspects, opinions, and sentiments. Key subtasks have evolved from basic ones like Aspect Extraction (AE), Aspect-Based Sentiment Classification (ABSC), and Aspect-Based Opinion Extraction (ABOE), to more complex joint extraction tasks involving triplets and quadruples, such as ACSTE (Aspect–Category–Sentiment Triplet Extraction), AOSTE (Aspect–Opinion–Sentiment Triplet Extraction), and ACOSQE (Aspect–Category–Opinion–Sentiment Quadruple Extraction). Notable contributions include the work of Yunsen Xian et al., who introduced a detailed sentiment extraction framework for Entity–Aspect– Opinion–Sentiment quadruples, and Hongjie Cai et al., who proposed leveraging implicit aspect/opinion components to improve the system's completeness and accuracy. Recent approaches have also explored semi-supervised learning with incomplete annotations, such as iACOS, and context-aware conditional tagging strategies, as in the work of Wang et al., enhancing the practical applicability of EAOS extraction in real-world scenarios.

In the Vietnamese context, although research in this domain remains limited, notable progress has been made. The UIT-ViSD4SA dataset, developed by the University of Information Technology – VNU-HCM, comprises over 11,000 user reviews annotated with ten distinct aspect categories. Additionally, datasets from the VLSP shared tasks in 2016 and 2018 have significantly contributed to the development of named entity recognition, text classification, and sentiment analysis systems for Vietnamese. On the modeling side, PhoBERT—a Vietnamese-specific variant of BERT—has demonstrated strong performance across various natural language processing tasks and is employed as the core encoder in the proposed EAOS architecture to fully exploit semantic information in Vietnamese texts. Real-Time Prediction Systems In recent years, a growing number of studies have explored the integration of machine learning techniques with big data processing platforms to build real-time prediction systems for social media data. Most of these systems focus on sentiment analysis, disease prediction, or anomaly detection based on streaming tweets or comments. For instance, Elzayady et al. (2018) utilized Apache Spark in combination with machine learning models for real-time sentiment analysis, optimizing preprocessing and model selection for improved accuracy. Ahmed et al. (2020) proposed a real-time system using Apache Spark and Apache Kafka to predict heart disease risk from tweets, evaluating multiple models including Decision Trees, SVM, and Random Forest, and applying grid search to select the optimal model. Zaki et al. (2020) developed a framework for collecting, processing, and visualizing Twitter data to analyze the real-time psychological state of Iraqi citizens, while Kilinc (2019) focused on fake account detection on Twitter using Spark MLlib. Another application by Kabir et al. (2020) involved tracking tweets from the United States during the COVID-19 pandemic to examine changes in topics, sentiment intensity, and subjectivity.

Although these studies demonstrate the effectiveness of combining Spark with machine learning for real-time social data analysis, most of them remain limited to overall sentiment classification and do not delve into finer-grained elements such as entities, aspects, or opinions. The proposed study extends this direction by building a real-time EAOS extraction system for Vietnamese YouTube comments. This system not only detects sentiment but also identifies and links textual spans representing entities, aspect categories, and opinions within each comment. By integrating deep learning models with big data frameworks such as Apache Spark, the system aims to enable fine-grained, real-time sentiment analysis from continuously updated social media streams, offering a novel direction for multidimensional opinion mining in dynamic online environments.

3 Real-Time EAOS Extraction System

The proposed real-time EAOS (Entity–Aspect–Opinion–Sentiment) extraction system consists of two main components: (i) an offline-trained EAOS extraction model and (ii) a streaming comment processing pipeline for applying the trained model to real-world data in real time. The overall system architecture is illustrated in Figure 1.

3.1 Offline Module: EAOS Extraction Model

Given the high velocity and volume of continuously generated social media data, traditional processing systems often fall short in terms of performance, scalability, and real-time responsiveness. This limitation is particularly evident in natural language processing tasks that require instant analysis and feedback upon data arrival. To address these challenges, the EAOS extraction system in this study is built upon Apache Spark, an open-source distributed computing framework that supports in-memory parallel processing and real-time data stream handling through the Spark Structured Streaming library.

Apache Spark is selected for its scalability, high processing speed, and rich ecosystem that supports numerous high-level libraries. In this system, Spark SQL is used for handling structured data, facilitating key text preprocessing steps such as filtering, normalization, and organization before feeding the data into the model. Structured Streaming is employed to construct a continuous data pipeline, ingesting real-time YouTube comments collected via the platform's official API. For deep learning model integration, PySpark, the Python API for Spark, enables efficient connection between the offline-trained PhoBERT model and the data stream being processed within Spark.

Specifically, incoming YouTube comments undergo a cleaning process that includes duplicate removal, text normalization, and the elimination of special characters and emojis. The cleaned data is then passed into the PhoBERT model to simultaneously predict four components: sentiment, aspect (aspect category), and the start and end positions of spans corresponding to entity and opinion expressions. These results are aggregated and visualized in real time, creating a continuous and automated loop from data collection to user feedback analysis on social media.

The use of Apache Spark not only ensures high-speed processing and scalability in the face of increasing data volume but also allows seamless integration with modern deep learning models. The combination of Spark's big data processing capabilities and PhoBERT's strong language representation power results in a stable, accurate, and efficient system tailored to the real-world characteristics of Vietnamese social media data.

3.2 Entity–Aspect–Opinion–Sentiment Extraction Module

3.2.1 Training Dataset

The dataset was constructed within the scope of this research to support the task of EAOS (Entity–Aspect–Opinion–Sentiment) extraction, where each data instance reflects one or more user opinions under popular entertainment videos on YouTube. Input data consists of user comments collected via Google Apps Script, which interfaces with the YouTube API to retrieve the latest comment threads. After collection, the data undergoes a preprocessing pipeline designed to remove noise and normalize the text. Specifically, the system eliminates non-linguistic elements such as hashtags, URLs, emojis, pictographic characters, and special symbols. The text is then converted to lowercase, punctuation is separated, duplicated characters are removed, and `vnCoreNLP` is applied for word segmentation and lexical normalization.

Manual annotation was carried out by a single annotator to maintain consistency, following detailed guidelines. Each comment may contain multiple EAOS quadruples, with each consisting of four components: entity, aspect (aspect category), opinion, and sentiment. Aspect categories are divided into five predefined groups: `APPEARANCE`, `CHARACTERISTIC`, `SPECIALIZE`, `TECHNICAL`, and `OTHER`. Sentiment is labeled as `positive`, `negative`, or `neutral`. For implicit components, the start–end index pair is assigned as $(-1, -1)$. The dataset is stored in tabular format, with each row containing the comment text,

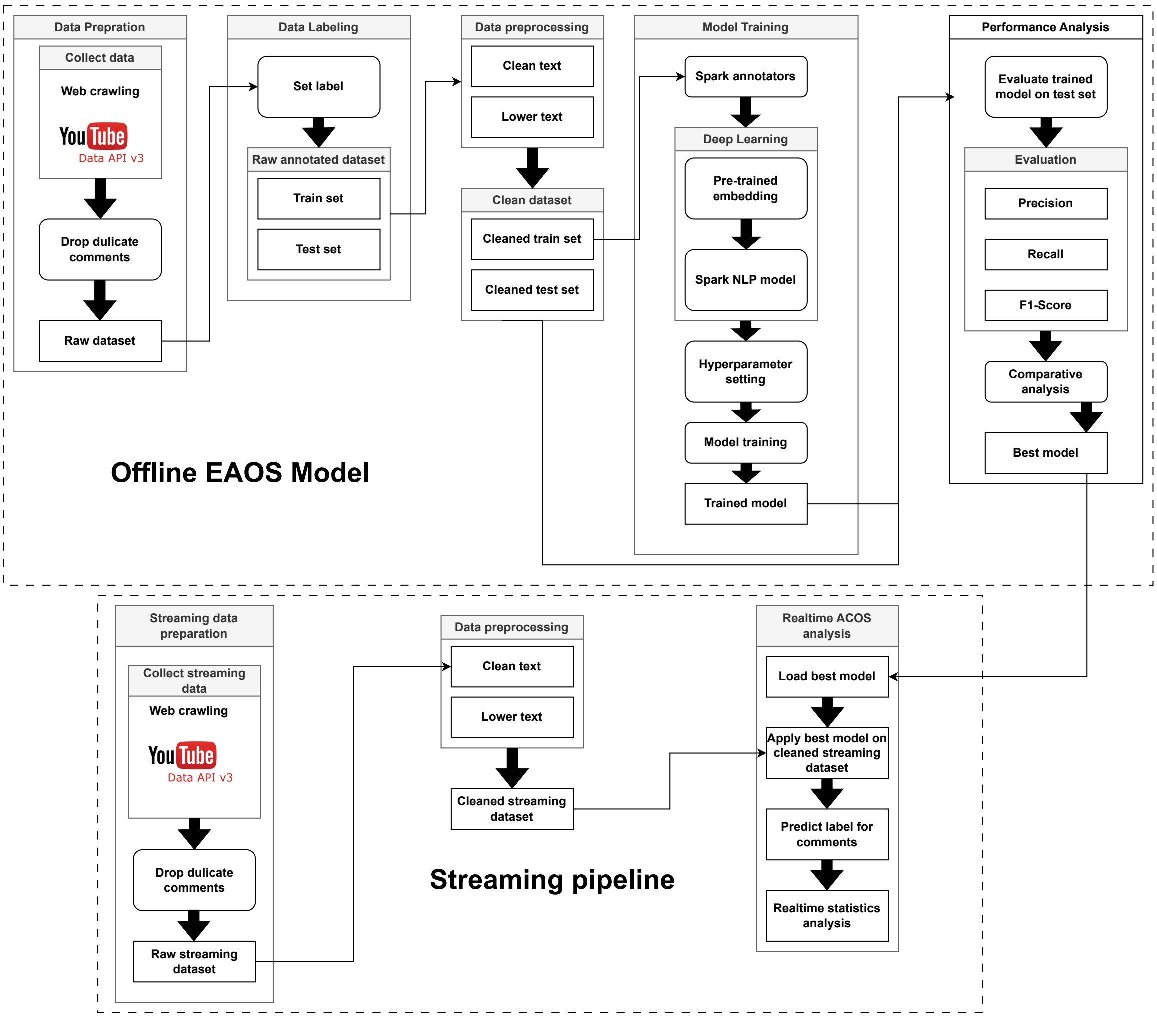


Figure 1: Overall EAOS extraction pipeline: offline training and real-time inference.

span positions, and the corresponding labels.

After annotation and data augmentation, the dataset contains a total of 10,065 comments, split into 9,065 training samples, 500 development samples, and 500 testing samples. Tables 1 and 2 summarize the distribution of sentiment and aspect categories in the training and development sets.

Table 1: Distribution of sentiment labels.

	Positive	Negative	Neutral
Train (9,065)	7,428	3,145	948
Dev (500)	402	142	69

Overall, the training set exhibits a distribution of 64.5% positive, 27.3% negative, and 8.2% neutral sentiment. Among the five aspect categories, “Specialize” accounts for the largest proportion, followed by “Other,” while the remaining three categories are more evenly distributed.

Table 2: Distribution of aspect categories.

Aspect Category	Train	Dev
Specialize	3,800	240
Other	2,907	194
Appearance	1,675	106
Technical	1,574	47
Characteristic	1,565	26

3.2.2 Preprocessing

Text preprocessing is essential, especially when dealing with social media data. The preprocessing pipeline removes emojis, mentions, hashtags, URLs, emails, and technical markers. Special characters, numbers, and extra whitespaces are also eliminated. The text is then lowercased, and `vnCoreNLP` is used for word segmentation and lemmatization.

Given the complexity of Vietnamese, which is monosyllabic and context-sensitive, tools like vnCoreNLP improve segmentation accuracy and grammatical preservation. Comments that are too short or too long are removed to ensure high-quality training data.

3.2.3 Deep Learning Models

LSTM and BiLSTM Baselines. LSTM and BiLSTM are used as baseline architectures for contextual learning. LSTM captures long-term dependencies but is limited to one-directional context. BiLSTM addresses this by processing input in both directions, which benefits tasks like EAOS where components may depend on future or past context.

EAOS Architecture. An advanced model combines PhoBERT, BiLSTM, Multi-Head Attention, and Graph Neural Networks (GNN). Input comments are encoded by PhoBERT, generating contextual vectors for each token. These vectors are passed through BiLSTM and Self-Attention layers.

The model performs three tasks: (1) extract candidate entities, (2) extract candidate opinions, (3) pair them into entity–opinion pairs. Both explicit and implicit expressions are captured. A GNN then models relationships between these pairs and predicts the aspect and sentiment labels.

The model outputs valid quadruples (e_i, a_j, o_k, s_l), where e_i is the entity, a_j the aspect, o_k the opinion, and s_l the sentiment. End-to-end training minimizes error propagation and improves overall accuracy.

This integrated model leverages PhoBERT for Vietnamese semantic encoding, BiLSTM for bidirectional context, Attention for focusing on important tokens, and GNN for relational modeling in EAOS extraction.

3.3 Online Module

3.3.1 Real-Time Comment Collection from YouTube

Google Apps Script is used in conjunction with the YouTube Data API to automatically collect comments from videos belonging to entertainment programs such as Rap Viet, 2 Days 1 Night, and The Masked Singer. The system periodically checks for new videos and downloads newly posted comments, removing duplicates to ensure data uniqueness. These comments are temporarily stored before the next processing stage. The retrieval process operates continuously via HTTP connections and employs OAuth functions for user authentication and secure data access.

3.3.2 Real-Time EAOS Analysis

After data collection, Spark Structured Streaming is used to process and stream comments to the pre-trained model from the offline phase. Each comment undergoes a preprocessing pipeline involving removal of special characters, emojis, and noise terms, normalization via vnCoreNLP, word segmentation, and standard text formatting. The processed text is then tokenized using the PhoBERT tokenizer and transformed into feature vectors.

These vectors are fed into the EAOS model, which may include components such as PhoBERT, BiLSTM, Attention, or GNN, depending on the selected architecture. The model simultaneously predicts: (i) entity and opinion spans, (ii) aspect category labels for each entity, and (iii) sentiment labels indicating positive, negative, or neutral attitudes. The extracted quadruples (e_i, a_j, o_k, s_l) are aggregated and stored in JSON or database format for statistical analysis.

The system then performs real-time statistical analyses such as: comment counts by sentiment type, distribution of user-focused aspects, frequency of specific entities or opinions over time, and shifts in viewer sentiment. These insights offer a comprehensive view of audience feedback on entertainment content and can assist content producers and media managers in making informed decisions.

4 Evaluation and Experimental Results

4.1 Offline Evaluation

Evaluation Metrics. To assess the effectiveness of the EAOS extraction models, we adopt three standard metrics commonly used in classification and sequence labeling tasks: **Precision**, **Recall**, and **F1-score**. These metrics are crucial in tasks with label imbalance and demand precise identification of EAOS components.

- **Precision** measures the proportion of correct positive predictions out of all predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

  where TP (True Positive) is the number of correct positive predictions, and FP (False Positive) is the number of incorrect positive predictions.

- **Recall** measures the proportion of actual positives that are correctly predicted:

$$\text{Recall} = \frac{TP}{TP + FN} \qquad (2)$$

where FN (False Negative) is the number of missed positive instances.

- **F1-score** is the harmonic mean of precision and recall:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (3)$$

It balances the trade-off between precision and recall, and is particularly suitable for multi-task, imbalanced settings like EAOS.

These metrics are applied not only to classification components (e.g., Apest and Sentiment) but also to span detection components (Entity and Opinion), where token-level accuracy and span-level coverage are measured.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
LSTM	32.52%	82.64%	32.52%	18.91%
LSTM + Attention	44.97%	52.28%	44.97%	40.56%
BiLSTM	53.14%	52.45%	53.14%	50.27%
BiLSTM + Attention	56.46%	57.76%	56.46%	53.26%
EAOS (Proposed)	**68.14%**	**69.23%**	**68.14%**	**68.28%**

Table 3: Experimental results comparing different models.

The table above presents the performance of different model architectures. The basic LSTM model achieves the lowest accuracy at 32.52%, despite a high precision of 82.64%, indicating a highly conservative prediction style that misses many true cases. Introducing Attention significantly improves recall (from 32.52% to 44.97%) and nearly doubles the F1-score to 40.56%.

The BiLSTM model, benefiting from bidirectional context encoding, shows better recall and F1-score (50.27%). Adding Attention further boosts performance to an accuracy of 56.46% and F1-score of 53.26%, marking the best among RNN-based models.

The final EAOS model, combining PhoBERT, BiLSTM, Attention, and GNN, achieves superior results across all metrics: accuracy of 68.14%, precision of 69.23%, recall of 68.14%, and an F1-score of 68.28%. This confirms the effectiveness of integrating contextualized language models and graph-based relational reasoning for comprehensive EAOS extraction.

In conclusion, the proposed EAOS model demonstrates substantial improvement over traditional architectures, validating the value of combining powerful representation layers, attention mechanisms, and graph structures in complex information extraction tasks.

4.2 Online Evaluation

Real-time Data Collection Speed The system collects comments from YouTube using the platform's API. During deployment, it takes approximately 5–7 minutes to process and store every 100 comments, depending on comment length, API response speed, and real-time data availability. This speed ensures timely data flow for real-time analysis.

EAOS Real-time Statistics From a total of 1,288 EAOS quads extracted in real-time, we compute the distribution of categories under different sentiment labels as shown in Table 2.

Sentiment	**Appearance**	**Characteristic**	**Specialize**	**Technical**	**Other**	**Total**
Negative	3	15	37	13	98	166
Positive	60	34	153	4	658	909
Neutral	0	7	2	4	200	213
Total	63	56	192	21	956	1288

Table 4: Category distribution by sentiment in real-time EAOS extraction

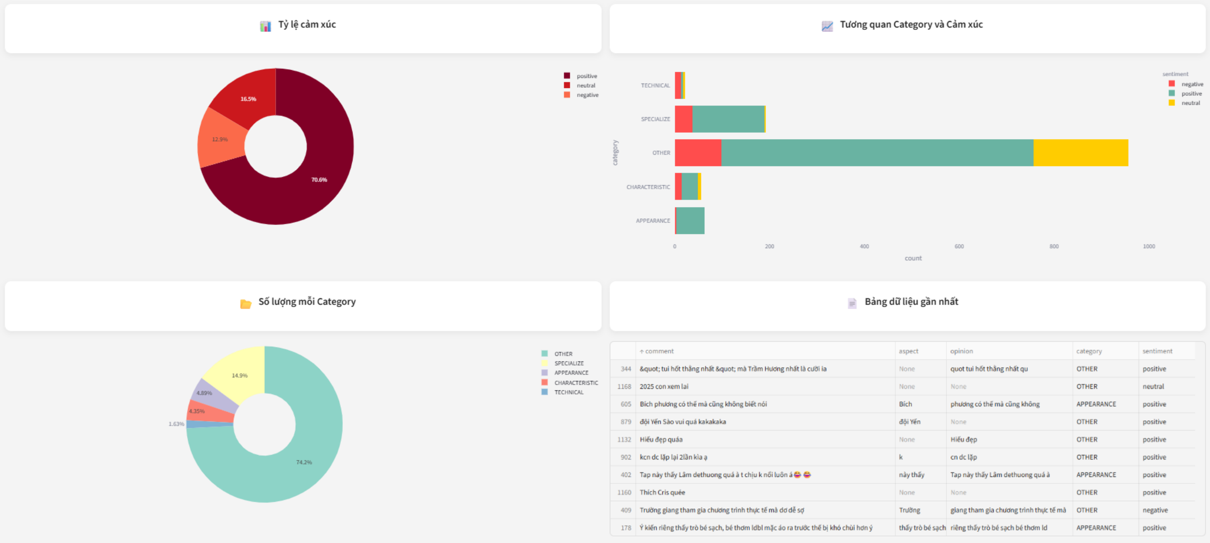
As shown, positive sentiment dominates (909 out of 1,288 EAOS quads, or 70.6%), which aligns with the nature of entertainment content on YouTube. "Other" and "Specialize" are the most frequent categories, indicating vague or skill-related user feedback.

Negative sentiment comments often focus on "Other" and "Specialize", suggesting viewer dissatisfaction tends to center on performance or less-defined categories. Neutral sentiment is mostly associated with the "Other" category, indicating descriptive or unclear expressions.

Overall, these statistics confirm the system's ability to reflect user reactions and trend shifts in real-time, supporting media monitoring and decision-making based on public feedback.

4.2.1 Error Analysis on Aspect Categorization

One notable observation is that a relatively high proportion of aspect labels fall into the `OTHER` category (approximately 25% of all annotations). Upon closer inspection, we found that many of these cases involve comments that are vague, multi-faceted, or outside the predefined set of four concrete categories (`APPEARANCE`, `CHARACTERISTIC`,



	↑ comment	aspect	opinion	category	sentiment
344	&quot; tui hốt thắng nhất &quot; mà Trầm Hương nhất là cười ia	None	quot tui hốt thắng nhất qu	OTHER	positive
1168	2025 con xem lai	None	None	OTHER	neutral
605	Bích phương có thể mà cũng không biết nói	Bích	phương có thể mà cũng không	APPEARANCE	positive
879	đội Yến Sào vui quá kakakaka	đội Yến	None	OTHER	positive
1132	Hiếu đẹp quáa	None	Hiếu đẹp	OTHER	positive
902	kcn dc lặp lại 2lần kìa ạ	k	cn dc lặp	OTHER	positive
402	Tap này thấy Lâm dethuong quá à t chịu k nổi luôn á😂 😂	này thấy	Tap này thấy Lâm dethuong quá à	APPEARANCE	positive
1160	Thích Cris quée	None	None	OTHER	positive
409	Trường giang tham gia chương trình thực tế mà dơ dễ sợ	Trường	giang tham gia chương trình thực tế mà	OTHER	negative
178	Ý kiến riêng thấy trò bé sạch, bé thơm ldbl mặc áo ra trước thế bị khó chùi hơn ý	thấy trò bé sạch	riêng thấy trò bé sạch bé thơm ld	APPEARANCE	positive

Figure 2: Real-time EAOS system dashboard illustrating sentiment distribution, category frequency, the correlation between sentiment and category, and a table of the most recent extracted data.

SPECIALIZE, TECHNICAL). For example, general praise such as ‘Hay quá!” (So good!”) or expressions of emotion like Thích quá tri” (Really love it”) do not explicitly refer to an identifiable aspect and are therefore placed in OTHER. This phenomenon reduces the usefulness of the system for fine-grained analysis. In future work, we plan to refine the annotation guidelines, introduce sub-categories within OTHER, and perform additional manual annotation to reduce the proportion of ambiguous cases. Such refinements are expected to improve both the interpretability and the practical value of the aspect extraction component.

5 Conclusion and Future Work

This work proposed and implemented a pioneering real-time information extraction system for Vietnamese social media comments, addressing the EAOS (Entity–Aspect–Opinion–Sentiment) task in a unified multi-task framework. To the best of our knowledge, this is the first system for Vietnamese that simultaneously detects entities, aspect categories, opinion spans, and sentiment within a single architecture—contrasting with previous studies that typically focused on individual subtasks such as sentiment analysis or aspect-based sentiment analysis. The system comprises two main components: an offline module for model training and an online module for large-scale, real-time EAOS extraction.

The proposed model leverages PhoBERT’s contextual semantic representations combined with task-specific linear layers for category classification, sentiment classification, and span extraction. The dataset—collected from popular entertainment programs, normalized, and manually annotated in the EAOS format—enabled effective fine-tuning of the model. Evaluation results showed that the proposed EAOS model outperformed strong baselines (LSTM, BiLSTM, Attention-based models), achieving an accuracy of 68.14% and an F1-score of 68.28%. These findings confirm the advantages of a transformer-based multi-task architecture for fine-grained sentiment analysis in Vietnamese.

The online component collects new user comments from YouTube via API and processes them using an Apache Spark Streaming–based pipeline, enabling large-scale, continuous data streams to be analyzed in real time. Experiments demonstrated an average latency of 50–60 seconds from comment posting to analysis completion. The system also provides real-time dashboards with statistics on aspects, sentiments, and opinions, offering actionable insights for marketing, brand tracking, and community feedback monitoring. With its flexibility and scalability, the architecture can be extended to other platforms such as Facebook, TikTok, and online forums, serving as a practical tool for decision support.

Limitations and Future Work — Despite these promising results, several limitations remain. First, the dataset is currently restricted to the entertainment domain, which may affect generalizability to other domains such as product reviews or news comments. Second, a relatively high proportion of aspects were labeled as "OTHER", suggesting the need for finer-grained annotation and further error analysis. Third, annotation was performed by a single annotator, which, although consistent, does not capture inter-annotator agreement. Finally, large language models (LLMs) were not considered in this study due to computational constraints; integrating such models represents an important direction for future work.

Future extensions will therefore focus on (i) expanding the dataset to multiple domains and ensuring higher annotation reliability, (ii) adapting the system to additional social media platforms, (iii) integrating it into real-time monitoring applications such as brand tracking and public opinion analysis, and (iv) exploring multilingual and cross-lingual settings. These directions aim to enhance the system's robustness, coverage, and applicability in real-world scenarios.

Acknowledgement

This research is funded by University of Information Technology - Vietnam National University Ho Chi Minh City under grant number D4-2025-14.

References

[1] Hongjie Cai, Rui Xia, Jianfei Yu. Aspect–Category–Opinion–Sentiment Quadruple Extraction with Implicit Aspects and Opinions. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (ACL 2021), pp. 340–350. 10.18653/v1/2021.acl-long.29

[2] Dan Ma, Jun Xu, Zongyu Wang, Xuezhi Cao, Yunsen Xian. Entity–Aspect–Opinion–Sentiment Quadruple Extraction for Fine-grained Sentiment Analysis. *arXiv preprint* arXiv:2311.16678 (2023).

[3] Wenya Zhang, Xin Li, Yang Deng, Lidong Bing, Wai Lam. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *arXiv preprint* arXiv:2203.01054 (2022).

[4] Soni S., Rambola A. A Comprehensive Review of ACOSQE with Implicit Aspects and Opinions. *Artificial Intelligence Review* (2023). 10.1007/s10462-023-10633-x

[5] Loris Di Quilio, Fabio Fioravanti. Evaluating the Aspect–Category–Opinion–Sentiment Analysis Task on a Custom Dataset. In: *CEUR Workshop Proceedings* (2022).

[6] Dang Van Thin, Lac Si Le, Vu Xuan Hoang, Ngan Luu-Thuy Nguyen. Investigating Monolingual and Multilingual BERT Models for Vietnamese Aspect Category Detection. *arXiv preprint* arXiv:2103.09519 (2021).

[7] Dat Quoc Nguyen, Anh Tuan Nguyen. PhoBERT: Pre-trained Language Models for Vietnamese. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1037–1042. 10.18653/v1/2020.findings-emnlp.92

[8] Hieu Nguyen, Linh Le, et al. ViSoBERT: A Pre-trained Language Model for Vietnamese Social Media Text Processing. *arXiv preprint* arXiv:2310.11166 (2023).

[9] Omar Elzayady, Xue Li, Mohamed Farouk. Real-Time Sentiment Analysis on Twitter Data Streams using Apache Spark. *International Journal of Data Science*, vol. 5, pp. 45–60 (2018).

[10] Mohammed Ahmed, Sara Al Murl. Real-time Heart Disease Risk Prediction from Twitter Data Using Apache Spark and Kafka. *Health Informatics Journal*, vol. 26, no. 2, pp. 1012–1028 (2020).

[11] Md. Kabir, Jane Smith. COVID-19 Twitter Sentiment Analysis using Apache Spark Streaming. In: *Proceedings of the 8th International Conference on Big Data Analytics* (2020).

[12] Nguyen Luong Chi, Nguyen Thi Minh Huyen, Nguyen Cam Tu, Le Hong Phuong. Vietnamese Open Information Extraction. *arXiv preprint* arXiv:1801.07804 (2018).