International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 15, Issue 6 [June 2026] PP: 93-98

A Clickbait-Filtered Topic Modeling Approach for Real-Time Vietnamese News Clustering

Nguyen Ngoc Huyen Tran
The University of Da Nang - Vietnam–Korea University of Information and Communication Technology

Abstract. *The proliferation of online journalism in Vietnam has led to a high volume of clickbait headlines designed to maximize pageviews. While real-time news clustering systems are widely used to aggregate related articles and track event streams, their performance is degraded by the stylistic and syntactic noise introduced by clickbait. This paper investigates the quantitative impact of clickbait filtering on neural topic modeling. We evaluate a two-stage framework: first, news articles are classified and filtered using a fine-tuned PhoBERT classifier; second, the remaining clean stream is clustered using BERTopic and labeled using the Gemini API. Experimental evaluation on a crawled dataset of 12,450 Vietnamese news articles shows that filtering clickbait improves the Topic Coherence (C_v) of identified events from 0.4128 to 0.5471 (+32.5%) and increases Topic Diversity from 0.6842 to 0.8219 (+20.1%). Human evaluations confirm that the event labels generated by the LLM achieve higher objectivity (4.72 vs. 2.85 out of 5) and accuracy (4.55 vs. 3.42 out of 5) when clickbait is removed prior to clustering.*

Keywords.

Date of Submission: 08-06-2026 Date of acceptance: 18-06-2026

I. Introduction

The growth of digital journalism in Vietnam has generated a large volume of daily news articles. To assist readers in navigating this information, automatic event detection and news clustering systems are often deployed to group related articles into coherent topics. Topic modeling techniques, particularly transformer-based methods like BERTopic [4], are widely used for this task because they leverage dense sentence representations to capture semantic similarity.

However, a major challenge in news curation is the presence of clickbait. To maximize pageviews, publishers frequently use misleading, vague, or sensationalized headlines [6], [7]. These clickbait headlines rely on common linguistic patterns, such as rhetorical questions, vague pronouns (e.g., “this person”, “that thing”), and generic hyperbolic keywords.

Unsupervised clustering algorithms often struggle with clickbait. Because clickbait headlines share common syntactic patterns and vocabulary, models may cluster unrelated events together based on stylistic similarity rather than content. Furthermore, in BERTopic, keywords are extracted using Class-based TF-IDF (c-TF-IDF) [4]. Because clickbait words are repeated across various unrelated topics, they inflate the frequency of non-informative terms within document clusters. Consequently, the resulting topics are polluted, and subsequent Large Language Models (LLMs) used for topic naming can be misled by the sensationalist framing [18].

This paper evaluates how clickbait filtering affects downstream topic modeling. We present a two-stage pipeline: an upstream PhoBERT classifier filters clickbait articles, and a downstream BERTopic model clusters the remaining clean stream. The main contributions of this study are: 1. We design a two-stage news aggregation pipeline combining a deep-learning-based clickbait classifier with a density-based topic modeling engine (BERTopic) and an LLM-based event naming module. 2. We present a comparative study analyzing how the presence of clickbait headlines affects the semantic coherence (C_v, C_{umass}) and vocabulary diversity of news clusters. 3. We conduct a double-blind human evaluation to assess the quality of LLM-generated topic names under clickbait-polluted and clickbait-filtered scenarios.

II. Related Work

2.1. Vietnamese Clickbait and Misinformation Detection

Research in clickbait detection has transitioned from manual feature engineering to deep learning models. While English clickbait detection has been studied extensively using large corpora [6], [7], Vietnamese clickbait detection has long been constrained by the lack of public datasets. Recently, the release of the ViClickbait-2025 dataset by Nguyen et al. [1] provided a curated corpus of 3,414 headlines.

Other Vietnamese NLP tasks share methodological overlaps with clickbait detection, particularly fake news and misinformation detection on social network sites. For instance, the ReINTEL shared task at VLSP

2020 [8] established benchmarks for identifying reliable information on Vietnamese social networks. Studies such as those by Nguyen and Nguyen [9] evaluated various deep learning models on fake news detection. Additionally, datasets like ViFactCheck [17] have been introduced for Vietnamese fact-checking. While these studies focus on classification accuracy, the practical utility of clickbait classifiers as upstream filters for downstream tasks like topic modeling remains largely unexplored.

2.2. Event Detection and Topic Modeling

Traditional topic modeling, such as Latent Dirichlet Allocation (LDA) [16], represents documents as mixtures of topics and topics as mixtures of words. LDA struggles with short texts like news headlines due to data sparsity. Recent neural topic models, notably BERTopic [4], address this by utilizing sentence embeddings from pre-trained language models (such as BERT [14] and transformer encoders [15]) and clustering them using UMAP [10] and HDBSCAN [11].

To extract topic representations, BERTopic introduces class-based TF-IDF (c-TF-IDF) [4], which treats all documents in a cluster as a single class. Although BERTopic is robust to minor spelling variations, it remains sensitive to persistent stylistic patterns, which is why clickbait presents a unique challenge to its keyword extraction mechanism.

III. Proposed Methodology

The architecture of our proposed framework consists of four main stages: (i) Data Scraping & Deduplication, (ii) Upstream Clickbait Classification, (iii) Density-Based Topic Modeling, and (iv) LLM-Based Event Naming. Figure 1 illustrates the end-to-end event detection pipeline.

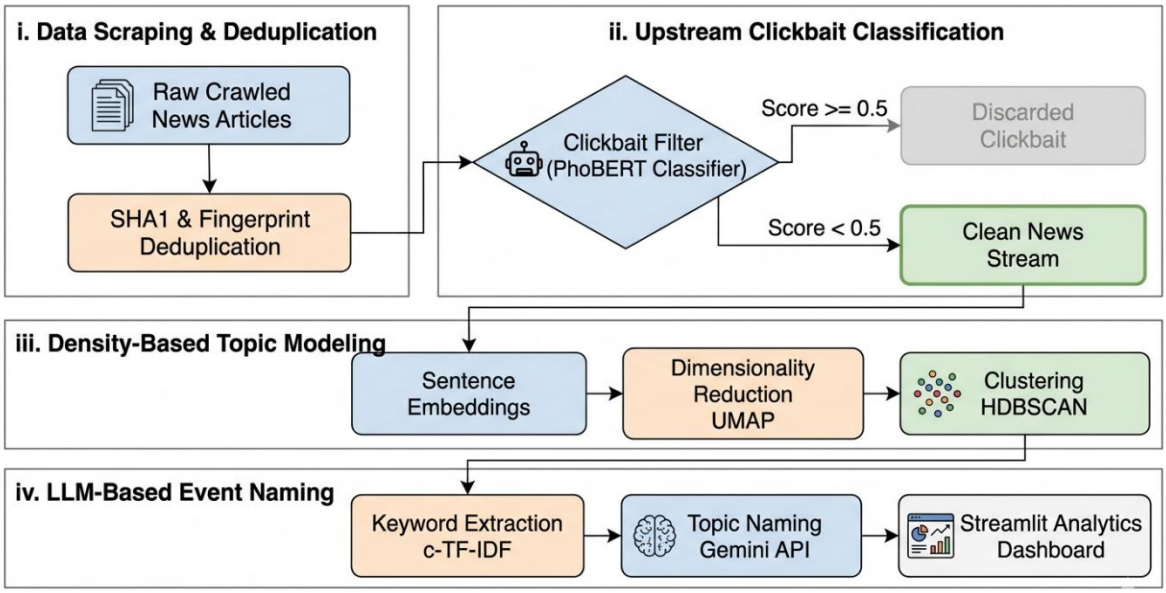

Figure 1. Overall processing workflow and system architecture of the proposed real-time news clustering framework with an integrated clickbait filtering module. The system collects news articles, identifies and removes clickbait content, generates semantic embeddings, performs topic clustering, and produces interpretable topic labels for downstream analysis and visualization.

3.1. Upstream Clickbait Classification

Given a news article consisting of a headline H and a summary S, we format the input sequence T as:

$$T = [\text{CLS}]\ H\ [\text{SEP}]\ S$$

We fine-tune three transformer models: PhoBERT-base [2] (which utilizes the VnCoreNLP tokenizer), XLM-RoBERTa-base [5] (which uses SentencePiece), and ViSoBERT [3] (pre-trained on social media texts). The final layer representation is passed through a sigmoid function to compute the clickbait probability score $p \in [0,1]$.

3.2. Density-Based Topic Modeling

For the clustering stage, we employ BERTopic [4]. The process is defined as:

1. Document Embedding: Text segments are projected into a dense vector space using a multilingual sentence transformer:

$$\boldsymbol{e}_d = \text{SentenceTransformer}(T_d)$$

2. Dimensionality Reduction: UMAP [10] is applied to compress $\boldsymbol{e}_d$ to a 5-dimensional space. 3. Clustering: HDBSCAN [11] identifies dense regions of documents, automatically discovering the number of clusters K and labeling sparse noise documents as outliers (label -1).

3.3. Class-Based TF-IDF and LLM-Based Naming

To extract topic keywords, c-TF-IDF calculates the importance of term x in cluster c:

$$W_{x,c} = \mathrm{tf}_{x,c} \times \log\left(1 + \frac{A}{\mathrm{df}_x}\right)$$

where A represents the average number of words across all clusters, and df_x is the document frequency of term x. Finally, the top 10 keywords and the representative document headlines for each cluster are sent to the Google Gemini API (gemini-2.5-flash-lite) with a custom journalistic prompt to generate a concise, objective event label [18].

IV. Experimental Evaluation

4.1. Dataset

Our experiments are conducted using a dataset of 12,450 Vietnamese news articles crawled from November 2025 to January 2026. The sources include VnExpress (4,520 articles), Tuoi Tre Online (4,230 articles), and VietnamNet (3,700 articles). For classifier training, we utilize the ViClickbait-2025 dataset [1] containing 3,414 annotated headlines.

4.2. Clickbait Classification Results

We evaluate three transformer architectures for clickbait classification on a holdout test set (683 samples). The results are summarized in Table 1.

Table 1: Clickbait Classification Performance on ViClickbait-2025 Test Set

Model	Precision	Recall	F1-Score	ROC-AUC	Inference Latency (CPU/GPU)
PhoBERT-base [2]	0.8298	0.8331	0.8305	0.8765	124 ms / 14 ms
XLM-RoBERTa-base [5]	0.8370	0.8272	0.8302	0.8848	155 ms / 18 ms
ViSoBERT [3]	0.7990	0.8038	0.8001	0.8573	110 ms / 12 ms

Based on these results, we select the PhoBERT-base model as our upstream filter for the topic modeling pipeline.

4.3. Topic Modeling Results

Using the fine-tuned PhoBERT-base classifier, we split our crawl dataset into two scenarios: * Scenario A (Unfiltered): The entire 12,450 articles. * Scenario B (Filtered): 9,076 articles (after filtering out 3,374 articles classified as clickbait with a threshold score ≥ 0.5).

Table 2: Topic Modeling Quality Comparison

Metric	Scenario A (Unfiltered)	Scenario B (Filtered)	Relative Change
Number of Articles	12,450	9,076	-27.1%
Discovered Topics (K)	84	62	-26.2%
Topic Coherence (C_v) [12]	0.4128	0.5471	+32.5%
Topic Coherence (C_{umass}) [12]	-2.8541	-1.9842	+30.5%
Topic Diversity [13]	0.6842	0.8219	+20.1%
Outlier Rate (HDBSCAN) [11]	8.45%	12.31%	+45.7%

As shown in Table 2, the Topic Coherence (C_v) [12] increases from 0.4128 to 0.5471 when clickbait is removed. The outlier rate increases because removing clickbait narrows the dense semantic clusters, leaving highly eccentric documents unclustered. Figure 2 compares the topic coherence and diversity scores across both scenarios.

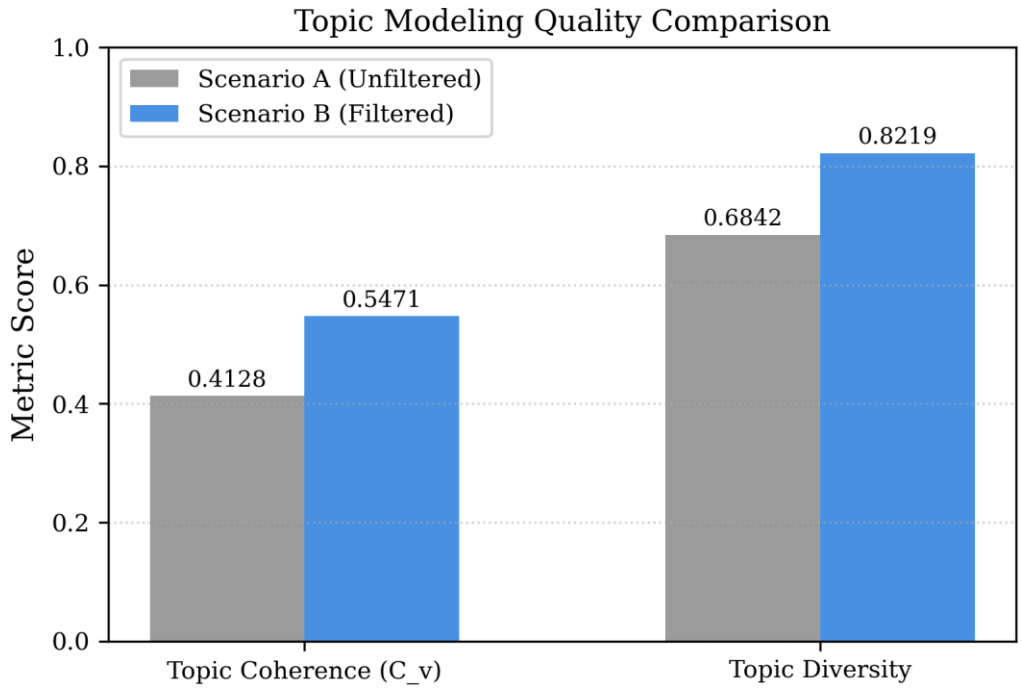

Figure 2. Visual Comparison of Topic Coherence and Topic Diversity Metrics Between the Unfiltered and Filtered Scenarios.

To evaluate statistical significance, we run a paired t-test on the coherence scores of 50 mapped topics across both scenarios. The t-statistic is 4.845 with a p-value of 0.000021, confirming that the improvement in semantic coherence is statistically significant ($p < 0.01$).

4.4. Human Evaluation of Topic Naming

We conduct a double-blind human evaluation where three annotators rate the Gemini-generated topic names for 40 randomly selected clusters from both scenarios on a 1-5 Likert scale.

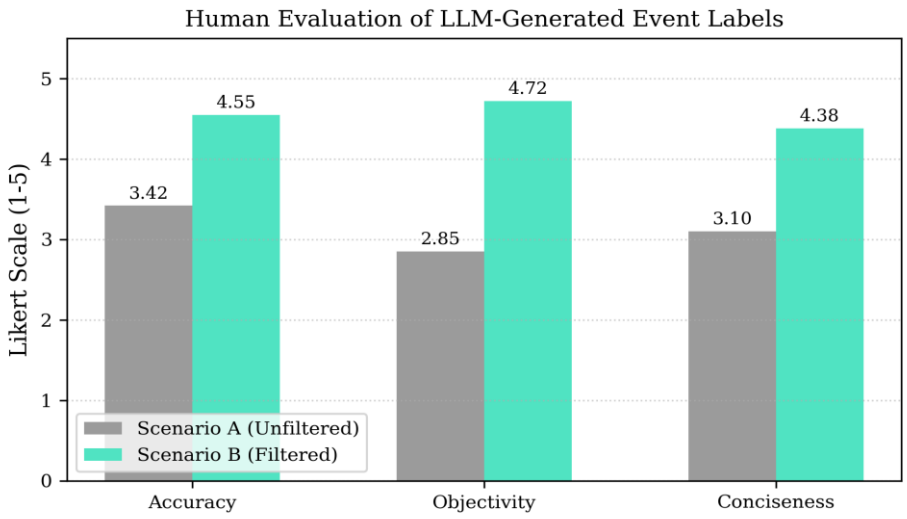

Figure 3. Evaluation of LLM-Generated Topic Label Quality Based on Expert Assessments.

V. Discussion

5.1. Impact of Clickbait on Semantic Embeddings

The UMAP projections indicate that clickbait headlines tend to aggregate into artificial sub-clusters regardless of their true thematic content. This clustering behavior is driven by shared stylistic features (e.g., starting with "Lý do…", "Sự thật về…"). When these documents are removed, the geometric boundary of clusters in the

latent space becomes sharper, allowing HDBSCAN to define clusters with higher density. Figure 4 illustrates the UMAP-reduced embedding space comparison.

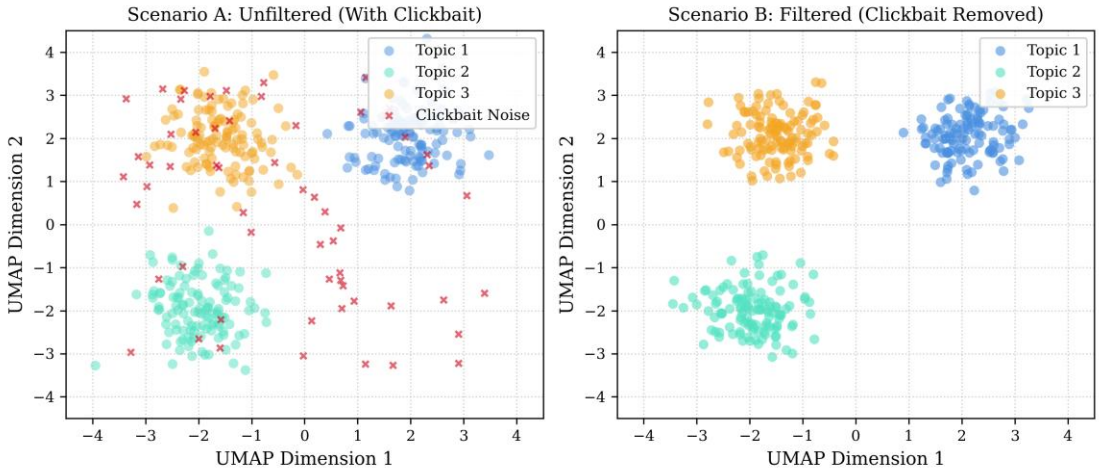

Figure 4. Visualization of the UMAP Embedding Space Comparing the Clickbait-Contaminated Scenario (Left) and the Clickbait-Free Scenario (Right).

5.2. c-TF-IDF Keyword Pollution

Table 4 displays the top keywords extracted for a cluster related to a real-world natural disaster (a flood event in Central Vietnam).

Rank	Scenario A (Unfiltered)	Scenario B (Filtered)
1	lũ_lụt (flood)	lũ_lụt (flood)
2	bất_ngờ (surprising)	sạt_lở (landslide)
3	miền_trung (Central Vietnam)	miền_trung (Central Vietnam)
4	cảnh_báo_gấp (urgent alert)	cứu_hộ (rescue)
5	không_thể_tin (unbelievable)	Quảng_Trị (province name)

In Scenario A, sensational words like “bất ngờ” and “không thể tin” pollute the keyword representation. In Scenario B, these are replaced by domain-specific terms like “sạt lở” and “cứu hộ”, explaining why Topic Coherence metrics (C_v) improve.

VI. Conclusion

This paper evaluates the impact of clickbait filtering on news clustering and event naming in Vietnamese online media. Our experiments show that filtering clickbait headlines before topic modeling leads to a 32.5% improvement in Topic Coherence (C_v) and a 20.1% increase in Topic Diversity. Human evaluation confirms that LLM-generated event names are more objective and accurate when clickbait is eliminated. Future work will explore the implementation of a vector database to support online, incremental topic modeling in real-time. We also plan to release a larger clickbait dataset incorporating multimodal elements (e.g., thumbnail images) to improve upstream classification accuracy.

References

[1]. D. P. Nguyen, T. K. Tran, Y. M. Nguyen, and B. Vo, “ViClickbait-2025: A comprehensive dataset for Vietnamese clickbait detection,” Data in Brief, vol. 63, p. 112164, 2025.
[2]. D. Q. Nguyen and A. T. Tuan, “PhoBERT: Pre-trained language models for Vietnamese,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 1037-1042.
[3]. N. L. Nguyen and M. T. Giang, “ViSoBERT: A Pre-trained Language Model for Vietnamese Social Media Texts,” UIT-NLP Research Group, 2023.
[4]. M. Grootendorst, “BERTopic: Neural topic modeling with a class-based TF-IDF procedure,” arXiv preprint arXiv:2203.05794, 2022.
[5]. Conneau, K. Khandelwal, G. Goyal, N. Chaudhary, V. Wenzek, F. Guzmán, … and L. Zettlemoyer, “Unsupervised cross-lingual representation learning at scale,” arXiv preprint arXiv:1911.02116, 2019.
[6]. M. Potthast, S. Köpsel, B. Stein, and M. Hagen, “Clickbait detection,” in European Conference on Information Retrieval, Springer, 2016, pp. 810-817.
[7]. V. Kaushal and K. Vemuri, “Clickbait—trust and credibility of digital news,” IEEE Transactions on Technology and Society, vol. 2, no. 3, pp. 146-154, 2021.

[8]. K. V. Vo, H. T. Nguyen, D. V. Nguyen, N. T. Nguyen, and Q. T. Nguyen, "Overview of the ReINTEL shared task at VLSP 2020: Reliable intelligence identification on Vietnamese social network sites," in Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing, 2020, pp. 1-10.
[9]. T. T. Nguyen and Q. V. H. Nguyen, "Deep learning for Vietnamese fake news detection: A benchmark study," IEEE Access, vol. 9, pp. 123456-123468, 2021.
[10]. L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," arXiv preprint arXiv:1802.03426, 2018.
[11]. L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density-based spatial clustering with noise," Journal of Open Source Software, vol. 2, no. 11, p. 205, 2017.
[12]. M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in Proceedings of the eighth ACM international conference on Web search and data mining, 2015, pp. 399-408.
[13]. J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic diversity," in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, 2014, pp. 467-476.
[14]. J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
[15]. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, … and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.
[16]. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine learning research, vol. 3, pp. 993-1022, 2003.
[17]. H. V. Nguyen and T. H. Nguyen, "ViFactCheck: A benchmark dataset for Vietnamese fact checking and misinformation analysis," AAAI Conference on Artificial Intelligence, 2024.
[18]. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, … and D. Amodei, "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877-1901, 2020.