Fact-checking for online advertisement posts

Tam T. Nguyen, Hao Nguyen Thi Phuong, Truong Phu Le, Binh T. Nguyen

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Fact-checking for online advertisement posts

Tam T. Nguyen[1,2,3], **Hao Nguyen Thi Phuong**[1,2,3], **Truong Phu Le**[1,2,3], **Binh T. Nguyen**[1,2,3*]
[1] *University of Science, Ho Chi Minh City, Vietnam*
[2] *Vietnam National University Ho Chi Minh City, Vietnam*
[3] *AISIA Research Lab, Vietnam*

Abstract

In Vietnam, the proliferation of social media has made these platforms prime targets for advertising. However, this trend has also led to a surge in misleading advertisements, particularly in the beauty and aesthetic sectors, posing significant health risks to consumers. Using data analysis techniques, this paper introduces the first methodological framework for detecting legal violations in beauty and aesthetic industry advertisements, such as false claims and unauthorized services. Additionally, we contribute a new dataset that we organized and collected, consisting of advertisement posts from beauty and aesthetic businesses on a social media platform, as well as their registered and approved information obtained from the government's public website. We evaluated our approach on this dataset and achieved reasonable and meaningful results, with accuracy, precision, and F1-score of 0.783, 0.686, and 0.703, respectively, using the BGEM3 model. The proposed solution aims to support regulatory agencies in identifying advertising violations and contribute to a safer and more transparent online environment for consumers.

1 Introduction

With over 72 million Facebook users in Vietnam as of January 2024 [1]. Social media is now an indispensable tool for advertising products and services. However, the increasing number of advertisements from beauty and aesthetic businesses on social media, a field that directly impacts health, has led to numerous cases of misleading or illegal information. However, verifying and validating this advertising information is a very challenging task [2], especially as the volume of information continues to grow. Therefore, research and development of technological solutions to support this mission are urgently needed [3].

Illegal advertisements in Vietnam include the following types: businesses without a license or with an expired license; businesses operating at locations different from those registered with government authorities; and advertisements containing false information and/or not aligned with the information registered and approved by regulatory agencies.

The contributions of this paper are as follows:

- Proposing an end-to-end methodological framework for detecting legal violations in beauty and aesthetic advertisements on social networks, with a primary focus on textual content.
- Introducing an integrated approach that combines linguistic feature extraction with a synthesized formula for matching textual semantic content between advertisement content and factual information.
- Presenting the first fact-checking dataset on beauty and aesthetic advertisements, which is used to evaluate our proposed framework. This dataset consists of 1,175 advertisement posts from beauty and aesthetic businesses collected from the Facebook platform, along with their corresponding registered information sourced from Ho Chi Minh City's government.

The paper can be organized as follows. In Section 2, we discuss related work and previous approaches. Section 3 describes our methodological

*Corresponding author: Binh T. Nguyen (e-mail: ngtbinh@hcmus.edu.vn).

[1] https://laodong.vn/y-te/hoat-dong-tham-my-da-phan-lam-sai-quang-cao-sai-1229568.ldo

[2] https://thanhnien.vn/giam-doc-so-y-te-tphcm-noi-3-thach-thuc-trong-quan-ly-tham-my-185230711162952828.htm

[3] https://nld.com.vn/bo-truong-nguyen-manh-hung-khong-the-dung-suc-nguoi-de-quan-ly-thuong-mai-dien-tu-196240605085505784.htm

framework and data workflow. Section 4 introduces our first fact-checking dataset on beauty and aesthetic advertisements. In Section 5, we detail our experimental setup and present the results with a thorough description. Finally, Section 6 outlines our conclusions and future work.

2 Related Work

The field of fake news detection and fact-checking has seen substantial global research, primarily on English-language datasets, using various processing techniques and models to analyze news and evidence. Monti et al. (2019) proposed a geometric deep learning model for fake news detection that leverages social network propagation patterns, generalizing classical CNNs to graph structures. Their method integrates diverse data types, achieving 92.7% ROC AUC accuracy and showcasing the benefits of propagation-based approaches over traditional content analysis.

Villela et al. (2023) conducted a systematic literature review on machine learning algorithms and datasets for fake news detection, identifying key algorithms like the Stacking Method, BiRNN, and CNN with accuracies of 99.9%, 99.8%, and 99.8%, respectively. Their research emphasizes the need for studies in real-time social network environments, addressing the limitations of controlled datasets. Sastrawan et al. (2022) explores fake news detection using deep learning methods, explicitly employing CNN, Bidirectional LSTM, and ResNet architectures. The study utilizes pre-trained word embeddings and trains on four datasets, incorporating data augmentation through back-translation to address class imbalances. Results indicate that the Bidirectional LSTM architecture consistently outperforms CNN and ResNet across all datasets.

Baarir and Djeffal (2021) developed a machine learning-based system for fake news detection, addressing challenges related to limited datasets and analysis techniques. They utilize the term frequency-inverse document frequency (TF-IDF) of the bag of words and n-grams for feature extraction, employing a Support Vector Machine (SVM) as the classifier. Their proposed dataset of fake and true news demonstrates the system's effectiveness.

In Vietnam, although fact-checking research based on Vietnamese datasets is still emerging, some notable studies have begun to appear: Hieu et al. (2020) presented a method for detecting fake news on Vietnamese social media platforms using an ensemble method combined with linguistic features extracted by PhoBERT. Their approach achieved an AUC score of 0.9521, ranking first on the test set at the 7th International Workshop on Vietnamese Language Processing and Pronunciation (VLSP). Pham et al. (2021) proposed a novel method for detecting fake news in Vietnamese by integrating the PhoBERT language model with Term Frequency-Inverse Document Frequency (TF-IDF) for vocabulary representation and a Convolutional Neural Network (CNN) for feature extraction. This model achieved an excellent AUC score of 0.9538 on raw data, trained and evaluated on the ReINTEL dataset.

Duong et al. (2022) proposed a model for fact-checking Vietnamese content by combining knowledge graphs (KG) with Bidirectional Encoder Representations from Transformers (BERT) deep learning techniques. This approach demonstrated high accuracy (up to 96%) on a Vietnamese dataset of 129,045 triples extracted from Wikipedia, enabling inference during fact-checking.

Tuan and Minh (2021) presents a method for fake news detection that combines textual features from a pre-trained BERT model with visual features from a VGG-19 model using a scale-dot product attention mechanism. Their approach achieves a 3.1% accuracy improvement over existing methods on a Twitter dataset, highlighting the effectiveness of multimodal feature fusion. Vo and Do (2023) developed a dataset of Vietnamese fake and factual news and evaluated deep learning models such as LSTM, bidirectional LSTM, and a CNN-bidirectional LSTM hybrid. Their study assessed model performance with metrics like AUC and highlighted the effectiveness of deep learning and neural network integration for Vietnamese fake news detection.

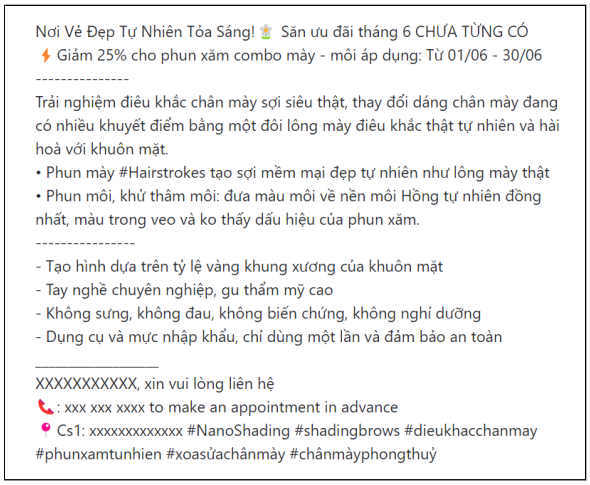
Most previous research on fake news detection in Vietnamese has not extensively explored fact-checking techniques, particularly for verifying advertising content on social media, which includes marketing-style information and data from regulatory agencies.

3 Methodology

3.1 Problem Formulation

Advertisement content typically comprises three formats: text, image, and video. This study focuses solely on the textual content, leaving the analysis of images and videos for future research. An adver-

tisement or post generally includes the following information: business name, address, phone number, license number, aesthetic techniques, promotional details, and other elements (such as emojis and hashtags), as shown in Figure 1.


Nơi Vẻ Đẹp Tự Nhiên Tỏa Sáng! Săn ưu đãi tháng 6 CHƯA TỪNG CÓ
Giảm 25% cho phun xăm combo mày - môi áp dụng: Từ 01/06 - 30/06

Trải nghiệm điêu khắc chân mày sợi siêu thật, thay đổi dáng chân mày đang có nhiều khuyết điểm bằng một đôi lông mày điêu khắc thật tự nhiên và hài hoà với khuôn mặt.
• Phun mày #Hairstrokes tạo sợi mềm mại đẹp tự nhiên như lông mày thật
• Phun môi, khử thâm môi: đưa màu môi về nền môi Hồng tự nhiên đồng nhất, màu trong veo và ko thấy dấu hiệu của phun xăm.

- Tạo hình dựa trên tỷ lệ vàng khung xương của khuôn mặt
- Tay nghề chuyên nghiệp, gu thẩm mỹ cao
- Không sưng, không đau, không biến chứng, không nghỉ dưỡng
- Dụng cụ và mực nhập khẩu, chỉ dùng một lần và đảm bảo an toàn
---
XXXXXXXXXX, xin vui lòng liên hệ
: xxx xxx xxxx to make an appointment in advance
Cs1: xxxxxxxxxxxx #NanoShading #shadingbrows #dieukhacchanmay #phunxamtunhien #xoasửachânmày #chânmàyphongthuỷ

Figure 1: A sample advertisement with summarized English translation as follows: 25% off on eyebrow and lip tattoo combo. Eyebrows: Microblading for a natural look. Lip Tattoo: Removes dark spots for a natural pink base. Pain-free, no swelling, and imported tools used.

To provide a tool for manually checking legal compliance, the government offers an official public website called the Information Search Portal for Healthcare Activities in Ho Chi Minh City (http://thongtin.medinet.org.vn). Consumers and regulatory agencies can use this portal to search for information on business names, license numbers, statuses, registered operating addresses, operational scopes, and permitted technical categories.

Although there may be discrepancies between the registered business name and the name used in advertisements (legally permissible discrepancies), the license number, registered address, operational scope, and aesthetic techniques listed in the advertisement must align with those provided on the government website.

The question arises as to how to extract and validate specific information from advertisement content to ensure compliance with the data available on the government website, as shown in Figure 2. The primary tasks are as follows: (1) Extract the license number and address from the advertisement text; (2) Verify that the extracted license number and address match the information on the government website. If the license number and address are valid, compare the technical categories in the adver-

```
"_source": {
  "name": [redacted],
  "address": [redacted],
  "permit": [redacted],
  "permit_issue_date": "2022-04-15",
  "status": "Đang hoạt động",
  "scopes": [
    {
      "scope_name": "Thực hiện kỹ thuật chuyên môn được Giám đốc
       Sở Y tế phê duyệt ban hành kèm theo Giấy phép hoạt động"
    }
  ],
  "registered_techniques": [
    {
      "technique_index": "53",
      "technique_name": "Đặt canuyn mũi hầu, miệng hầu",
      "technique_operators": [],
      "technique_subcategory": "HÔ HẤP",
      "technique_category": "HỒI SỨC CẤP CỨU VÀ CHỐNG ĐỘC"
    },
    {
      "technique_index": "65",
      "technique_name": "Bóp bóng Ambu qua mặt nạ",
      "technique_operators": [],
      "technique_subcategory": "HÔ HẤP",
      "technique_category": "HỒI SỨC CẤP CỨU VÀ CHỐNG ĐỘC"
    },
```


Figure 2: A sample of factual data containing information registered with the government shows that some licensed techniques include nasopharyngeal and oropharyngeal cannula insertion, ambu bag ventilation through a mask (belonging to the group of emergency resuscitation and detoxification). Status: Active, licensed on April 15, 2022.

tisement with those listed on the website. (3) Given that exact matches are unlikely due to variations in text representation, traditional text comparison methods or keyword-based approaches are insufficient. The challenge is to devise a method for text comparison that goes beyond exact text matching, allowing for a robust comparison of technical categories despite potential differences in phrasing or terminology.

3.2 Preliminary

Vietnamese SBERT (Vs-BERT) (Phan et al., 2022) is a sentence embedding model based on PhoBERT, optimized for Vietnamese. It improves NLP tasks such as text classification and sentence similarity, achieving 5-10% performance gains over traditional methods. Trained on a diverse dataset, it ensures strong generalization and can be easily integrated into NLP applications without the need for retraining.

PhoBERT (Nguyen and Nguyen, 2020) is a Vietnamese language model based on BERT. It is trained on a diverse dataset of newspapers, books, and web documents, capturing the unique linguistic features of Vietnamese. PhoBERT outperforms multilingual models like BERT and XLM-R in Vietnamese NLP tasks, including text classifica-

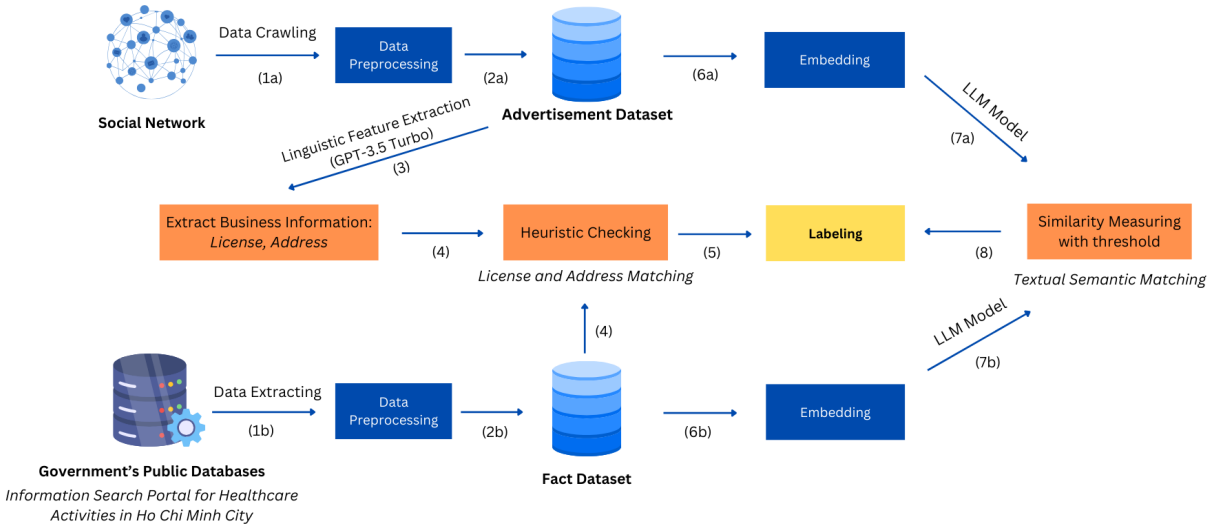


Figure 3: An end-to-end methodological framework for detecting illegal advertisements

tion, entity recognition, and sentiment analysis.

BGE-m3 (Chen et al., 2024) is an embedding model known for its multilingual capabilities, multifunctionality, and multi-granularity. It supports over 100 languages and excels in both multilingual and cross-lingual retrieval tasks. The model is capable of performing dense, multi-vector, and sparse retrieval for inputs ranging from short sentences to long documents, with a maximum length of 8192 tokens.

MiniLLM (Gu et al., 2023) is a Knowledge Distillation (KD) method for distilling LLMs into smaller language models. It addresses the limitations of previous KD methods for generative language models. MiniLLM produces more accurate responses with higher overall quality, lower exposure bias, better calibration, and improved long-text generation performance compared to baseline models.

GPT-3.5 Turbo (Gue et al., 2024) is an enhanced version of the GPT-3 model, designed to deliver higher performance and accuracy in text generation. Compared to previous versions, GPT-3.5 Turbo improves semantic understanding and contextual awareness, resulting in more accurate and natural responses across a wide range of scenarios. This model excels in handling long and complex texts while also reducing errors and enhancing response calibration and balance.

Cosine similarity (Rahutomo et al., 2012) measures the similarity between two vectors by calculating the cosine of the angle between them. A higher value indicates greater similarity, as the vectors point in similar directions. This metric is especially useful in text analysis for assessing document similarity, effectively addressing the limitations of Euclidean distance, which can be misleading for documents of varying lengths.

3.3 Our proposed solution

3.3.1 A methodological framework

In order to create a comprehensive system capable of effectively detecting violations, we propose a comprehensive methodological framework for identifying legal violations in beauty and aesthetic advertisements on social networks, with a primary focus on textual content, as shown in Figure 3.

Data Collection (1a & 1b): Data is gathered from social media platforms and official fact sources.

Data Processing (2a & 2b): We clean and remove unnecessary information from the data, then save it to the Advertisement dataset and the Fact dataset.

Extract Business information (3): For each advertisement, we use GPT-3.5 Turbo to extract the license number and address from the content for heuristic checking.

Heuristic Checking (4): The extracted license number and address are checked for accuracy against those from the Fact dataset.

Labeling (5): If either the license information or the registered address is incorrect compared to the official registration, the data is labeled as a violation.

Embedding (6a, 7a & 6b, 7b): If the license number and address match the registered information, proceed with embedding to prepare for

semantic content matching in the next step.

Technical content extraction (3): In each advertisement, we use GPT-3.5 Turbo to extract the technical categories from the content, and the model returns them as a list.

Similarity Measurement (8): Using an LLM model, the framework performs textual semantic matching to measure the similarity between datasets, as shown in Figure 4. A threshold-based similarity measure determines whether an advertisement is potentially in violation or not.

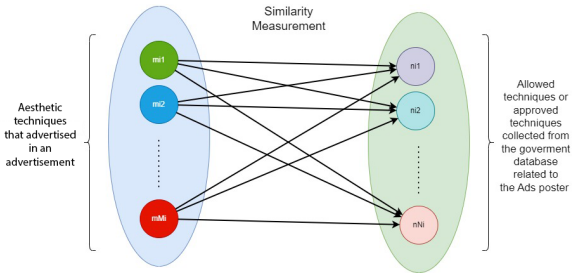


Figure 4: An illustration depicting the synthesized formula that we use

3.3.2 A synthesized formula for matching the textual semantic content of advertisement with factual information

Assume we have an advertisement post $\mathcal{M}$ containing m technical content items and a set $\mathcal{N}$ containing n registered technical content items. The objective is to examine the semantic similarity between two sets $\mathcal{M}$ and $\mathcal{N}$. To achieve this objective, it is necessary to calculate the similarity between each element in $\mathcal{M}$ and each element in $\mathcal{N}$.

Calculating Similarity

First, we calculate the similarity between each pair (m_i, n_j) where $m_i \in \mathcal{M}$ and $n_j \in \mathcal{N}$ using the cosine similarity function $\text{sim}(m_i, n_j)$. This will create a similarity matrix $\mathbf{S}$ of size $m \times n$, with each element s_{ij} defined as:

$$s_{ij} = \text{sim}(m_i, n_j)$$

Optimizing Similarity

For each content item m_i in $\mathcal{M}$, we calculate the maximum similarity value from the corresponding row in matrix $\mathbf{S}$:

$$\max_i = \max_{1 \le j \le n} s_{ij}$$

The result of this step is a vector $\mathbf{v}$ of size $m \times 1$, with each element $v_i = \max_i$.

Calculating Overall Similarity

Finally, to ensure that if even one technical content item in the advertisement has a low similarity score compared to the registered technical content, the entire advertisement will be considered to have low similarity (according to the principle that if one technical content violates, the entire advertisement is considered a violation), we calculate the minimum value of the vector $\mathbf{v}$:

$$\text{sim}_{\text{final}} = \min_{1 \le i \le m} v_i$$

The value $\text{sim}_{\text{final}}$ ranges from -1 to 1 and serves as the final measure of similarity between the advertisement post and the registered technical content.

Interpretation

If $\text{sim}_{\text{final}}$ is high (close to 1), this indicates that the content of the advertisement post is not in violation, meaning it closely matches at least one of the registered content items.

Conversely, if $\text{sim}_{\text{final}}$ is low (close to -1), this suggests that the advertisement post is likely to be in violation since none of its content is sufficiently similar to the registered items.

3.4 Performance Evaluation

To evaluate the performance of our proposed method, we utilize standard metrics commonly employed in information retrieval and natural language processing tasks. These metrics provide a comprehensive assessment of the method's ability:

Accuracy: This metric represents the proportion of data pairs correctly identified as either similar or dissimilar.

F1-score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the system's performance.

Precision: Precision measures the proportion of data pairs identified as similar that are similar. It indicates the system's ability to avoid false positives.

4 Datasets

4.1 Data Description

Our proposed dataset on the beauty and aesthetic sector comprises advertisement content collected from Facebook, along with corresponding registered information sourced from the Information Search Portal for Healthcare Activities in Ho Chi Minh City, as mentioned above. This dataset includes business licenses (mapped to operational

scopes and technical categories), technical licenses (mapped to permitted services), operating addresses, operational scopes, permitted technical categories, and aesthetic techniques extracted by GPT-3.5 Turbo from the advertisement content. The entire data processing process to create the proposed dataset is described in Figure 5.

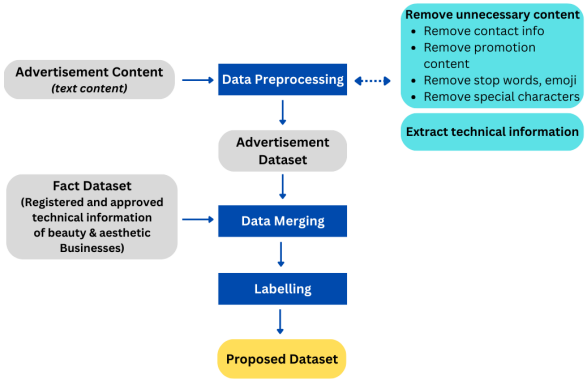


Figure 5: The data flow of the entire process of building the dataset, from collection and preprocessing to labeling the data

4.2 Data Acquisition

The first dataset includes 1,175 advertisements from 283 beauty and aesthetic businesses. Our data team manually collected this data from fan pages and groups on Facebook. It comprises 929 data points labeled as 0 (78.9%) and 246 labeled as 1 (21.1%), sourced from 30 aesthetic businesses (specialized clinics or hospitals) and 253 beauty businesses, with five businesses holding both business and technical licenses.

The second dataset contains registered technical categories from the Ho Chi Minh City Department of Health, including facility name, address, license, scope of practice, and techniques. It includes 9,164 specialized clinics or cosmetic hospitals and 3,890 beauty facilities, with technique lists from 916 specialized clinics. This data was sourced from the Information Search Portal for Healthcare Activities in Ho Chi Minh City: http://thongtin.medinet.org.vn.

Our dataset labeling process is derived from the two datasets previously described. A post is classified as a violation and assigned a label of 1 if it advertises technical procedures not listed among the categories licensed by the Ho Chi Minh City Department of Health. Conversely, if the advertised techniques are within the licensed categories, the post is labeled 0, indicating no violation.

Subsequently, we used the licensing information to map the above datasets into a unified dataset.

4.3 Data Preprocessing

Our dataset labeling process is derived from the two datasets previously described. A post is classified as a violation and assigned a label of 1 if it advertises technical procedures not listed among the categories licensed by the Ho Chi Minh City Department of Health. Conversely, if the advertised techniques are within the licensed categories, the post is labeled as 0, indicating no violation.

For the content, we perform the following preprocessing steps: convert uppercase letters to lowercase; remove emojis and non-alphabetic characters; build a set of Vietnamese stopwords tailored to the dataset; then remove these stopwords from the dataset. Finally, we use GPT-3.5 Turbo to extract technical categories from the content; an example of the extracted content results is shown in Table 1.

Meanwhile, the data in the beauty facility dataset, sourced from the Information Search Portal for Healthcare Activities in Ho Chi Minh City, is in JSON format. Therefore, the preprocessing of this dataset differs from the advertisement dataset described above, including the following steps: Convert JSON to CSV; Transform JSON format into a Python list for the technical categories and scope of activities.

4.4 Statistical Analysis

The dataset, obtained by merging two previously described datasets, comprises 1,175 records and nine variables related to beauty and aesthetic advertisements and their registration status with the authorities.

address: Contains 1,171 non-null entries indicating the location. The majority of entries are complete, with only four missing values.

license: This column is significantly sparse, with only 102 non-null entries, suggesting limited availability of specific licensing information.

business license: This variable has 1,095 non-null entries, providing the business licenses associated with the records. This column is relatively well-populated.

content: Contains textual data from advertisements with 1,174 non-null entries, making it nearly complete.

cleaned content: This variable is fully populated with 1,175 non-null entries and contains cleaned and processed textual data.

Before prompting	After prompting
[Vietnamese Version] mụn nỗi đau mụn nhiên sinh chẳng dưng đi mụn dạng viêm thâm rỗ cái dại trị mụn là từ mấy mụn ko sao chữa nghỉ chữa vân vân mây mây lý và ca nỗi khổ lẽ chăm sóc da kĩ chữa lẽ ko chủ quan mụn mọc khỏi hối hận quá đây giải quyết để làn da đẹp mỹ	Chăm sóc da kỹ lưỡng; Giải quyết mụn; Làm đẹp da
[English Version] Pimples, the pain of pimples, naturally appear out of nowhere, inflammatory pimples, scars, the foolishness of treating pimples comes from not knowing how to treat them, taking breaks from treatment, and so on and so forth. The reason and case for suffering are probably due to not taking proper care of the skin and thinking it's not serious. When pimples appear and are not treated, you will regret it. Here's how to resolve it to get beautiful skin.	Thorough skin care; Acne treatment; Skin beautification.
[Vietnamese Version] chương trình khuyến mãi nặn mụn 250k áp dụng masssage cổ vai gáy nặn mụn nắng nè	Khuyến mãi; Massage cổ vai gáy; Nặn mụn
[English Version] Promotion program for acne treatment at 250k includes neck and shoulder massage, acne treatment, and sun protection.	Promotion; Neck and shoulder massage; Acne extraction

Table 1: The output after applying GPT-3.5 Turbo for extracting technical categories from advertisement content in English and Vietnamese.

scopes: Only 98 records have non-null values in this column, reflecting the specific scopes of the services advertised.

allowed serviced: 1,094 entries are non-null, detailing the officially permitted services.

tech list: Contains data on specific techniques mentioned in the advertisements, but with only 53 non-null entries, this column is sparsely populated.

scopes tech list: With 1,170 non-null entries, this variable combines the scopes, techniques list, and allowed services to provide comprehensive technical categories, making this a crucial variable for identifying potential violations.

5 Experiments

5.1 Experimental Settings

The experiments were conducted using Python 3.8 on the Google Colab CPU environment. In the initial experimental phase, we evaluated content matching between social network posts and factual documents using embeddings generated by MiniLLM, BartPho, BGEM3, Vietnamese S-BERT, and PhoBERT models. The cosine similarity metric was used to measure the similarity between the text embeddings. The performance of the text embedding models was evaluated using different similarity thresholds, ranging from 0.1 to 0.9 in increments of 0.1. This was done to determine the optimal threshold for classifying text documents as similar or dissimilar based on their embeddings. Accuracy, Precision, and F1-score are the metrics used to assess the performance of the text embedding models.

5.2 Results

According to the metrics - accuracy, precision, and F1 score - summarized in Table 2, the models assessed include PhoBert, BartPHO, and MiniLLM, all of which exhibited similar performance with accuracy consistently around 79% across various thresholds. The prevalence of this class imbalance suggests that these models may not be effectively learning to discriminate between classes. Our analysis further identifies BGEM3 and Vietnamese S-BERT as the most effective large language models in this study, as shown in Figure 6. While both models demonstrate high accuracy, especially within thresholds ranging from 0.1 to 0.3, the minimal variation in their results suggests they may be inclined to predict predominantly zero la-

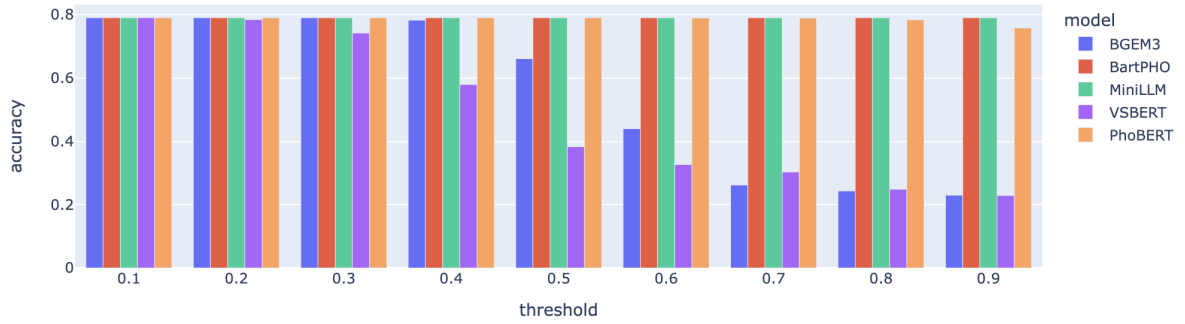


Figure 6: The accuracy performance among different models.

Accuracy	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
BGEM3	0.791	0.791	0.791	**0.783**	0.661	0.440	0.262	0.243	0.230
PhoBERT	0.791	0.791	0.791	0.791	0.791	0.790	0.790	0.784	0.758
VSBERT	0.791	0.785	0.742	0.580	0.383	0.327	0.303	0.249	0.229
MiniLLM	0.791	0.791	0.791	0.791	0.791	0.791	0.791	0.791	0.791
BartPHO	0.791	0.791	0.791	0.791	0.791	0.791	0.791	0.791	0.791
Precision	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
BGEM3	0.625	0.625	0.625	0.686	0.667	0.719	0.675	0.703	0.676
PhoBERT	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.684	0.662
VSBERT	0.625	0.624	0.663	0.684	0.732	0.768	0.772	0.724	0.672
MiniLLM	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625
BartPHO	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625
F1-score	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
BGEM3	0.698	0.698	0.698	0.703	0.664	0.472	0.199	0.153	0.124
PhoBERT	0.698	0.698	0.698	0.698	0.698	0.698	0.698	0.702	0.696
VSBERT	0.698	0.695	0.693	0.616	0.393	0.298	0.257	0.161	0.123
MiniLLM	0.698	0.698	0.698	0.698	0.698	0.698	0.698	0.698	0.698
BartPHO	0.698	0.698	0.698	0.698	0.698	0.698	0.698	0.698	0.698

Table 2: The performance of different methods according to the threshold from 0.1 to 0.9

bels for the dataset. At a threshold of 0.4, BGEM3 achieved an accuracy of 0.783 and an F1 score of 0.703, indicating a strong performance in classifying infringing social media posts. In contrast, Vietnamese S-BERT recorded an accuracy of 0.58 and an F1 score of 0.616. Overall, based on the data presented and the performance of the large language models assessed, we conclude that BGEM3 and Vietnamese Sbert exhibit significant potential for accurately matching and classifying infringing social media posts, particularly when utilizing binary labels of 0 and 1.

6 Conclusion

We have proposed an approach to fact-checking Vietnamese advertisement posts in the beauty and aesthetic sector. We introduced an end-to-end framework and a method for similarity calculations in the fact-checking process. Additionally, we created a new dataset to support further research. Our results show that two out of five large language models are effective in this context.

We plan to gather more data from Facebook and other social media platforms, such as images and videos, to enhance our data collection and improve our results. We also aim to incorporate images and videos from advertisements for semantic matching. Implementing a comprehensive solution based on the framework proposed in this paper will effectively validate advertising information on social networks, benefiting regulatory agencies and consumers.

References

Nihel Fatima Baarir and Abdelhamid Djeffal. 2021. Fake news detection using machine learning. In *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*, pages 125–130.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Huong To Duong, Van Hai Do, and Phuc Doa. 2022. Vietnamese fact checking based on the knowledge graph and deep learning.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Minillm: Knowledge distillation of large language models.

Celeste Ci Ying Gue, Noorul Dharajath Abdul Rahim, William Rojas-Carabali, Rupesh Agrawal, Palvannan RK, John Abisheganaden, and Wan Fen Yip. 2024. Evaluating the openai's gpt-3.5 turbo's performance in extracting information from scientific articles on diabetic retinopathy.

Thuan Nguyen Hieu, Hieu Cao Nguyen Minh, Hung To Van, and Bang Vo Quoc. 2020. ReINTEL challenge 2020: Vietnamese fake news detection using Ensemble model with PhoBERT embeddings. In *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*, Hanoi, Vietnam. Association for Computational Linguistics.

Federico Monti, Fabrizio Frasca, Davide Eynard, and Michael M. Bronstein Damon Mannion. 2019. Fake news detection on social media using geometric deep learning.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese.

Ngoc Dong Pham, Thi Hanh Le, Thanh Dat Do, Thanh Toan Vuong, Thi Hong Vuong, and Quan Thuy Ha. 2021. Vietnamese fake news detection based on hybrid transfer learning model and tf-idf.

Quoc Long Phan, Tran Huu Phuoc Doan, Ngoc Hieu Le, Duy Tran, and Tuong Nguyen Huynh. 2022. Vietnamese sentence paraphrase identification using sentence-bert and phobert.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity.

Kadek Sastrawan, I.P.A. Bayupati, and Dewa Made Sri Arsa. 2022. Detection of fake news using deep learning cnn–rnn based methods.

Nguyen Manh Duc Tuan and Pham Quang Nhat Minh. 2021. Multimodal fusion with BERT and attention mechanism for fake news detection. *CoRR*, abs/2104.11476.

Humberto Fernandes Villela, Jurema Suely de Araújo Nery Ribeiro Fábio Corrêa, Air Rabelo, and Dárlinton Barbosa Feres Carvalho. 2023. Fake news detection: a systematic literature review of machine learning algorithms and datasets.

Duc Vinh Vo and Phuc Do. 2023. Detecting vietnamese fake news. *CTU Journal of Innovation and Sustainable Development*, 15(Special issue: ISDS):39–46.