A Social Listening System for Beauty Products Using Aspect-Based Sentiment Analysis

Thanh-Nhi Nguyen, Trong-Hop Do

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

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

A Social Listening System for Beauty Products Using Aspect-Based Sentiment Analysis

Thanh-Nhi Nguyen[1,2], **Trong-Hop Do**[1,2]

[1]University of Information Technology,
[2]Vietnam National University Ho Chi Minh City
Corresponding author: Trong-Hop Do (hopdt@uit.edu.vn)

Abstract

In the digital age, the vast amount of unstructured data from online platforms poses significant challenges for businesses aiming to extract actionable insights. Social listening, empowered by big data analytics, emerges as a vital tool to monitor and understand consumer sentiments and trends. This paper introduces a specialized social listening system tailored for the beauty industry, integrating textual and video data sources. Central to our approach is Aspect-Based Sentiment Analysis (ABSA), which dissects consumer feedback into specific product aspects to discern nuanced sentiments. We present novel methodologies for data normalization and subject identification, crucial for enhancing the granularity and accuracy of sentiment analysis in this domain. Normalization here refers to both lexical normalization of noisy social media text and post-ASR transcript normalization, ensuring consistent and comparable linguistic input across modalities. Experimental results demonstrate the effectiveness of our system in extracting and classifying sentiments related to beauty products, surpassing existing benchmark.

1 Introduction

In today's digital age, the vast amounts of unstructured data generated from online sources, such as social media platforms, blogs, forums, and e-commerce reviews, present a significant challenge for businesses and researchers alike. These platforms are replete with valuable insights that can inform strategic decisions, yet the sheer volume and complexity of the data make it difficult to analyze and utilize effectively. One promising solution to this challenge is social listening powered by big data analytics.

Social listening is defined as the process of monitoring digital conversations to understand what customers are saying about a brand, product, or industry online. The term "social listening" refers to the practice of monitoring and analyzing user-generated content on social media platforms (Westermann and Forthmann, 2021). This involves not just tracking mentions and comments but also analyzing the sentiment, context, and trends within these conversations. By leveraging big data technologies, social listening can transform vast amounts of unstructured data into actionable insights, helping businesses to stay ahead of market trends, understand customer needs, and improve their products and services.

Numerous studies have demonstrated the benefits of social listening. For instance, it has been shown to enhance customer relationship management, support product development, and improve marketing strategies (Moe and Schweidel, 2017). By tapping into the collective voice of consumers, companies can gain a deeper understanding of market dynamics and consumer preferences.

In Vietnam, the beauty industry is particularly notable for its dynamic growth and consumer engagement. According to recent metrics, the beauty sector leads the market in sales and has experienced growth rates of nearly 150% during 2021–2023, according to recent e-commerce metrics [1]. This remarkable growth highlights the importance of understanding consumer behavior and trends within this sector. Given the high engagement levels and the substantial market share of beauty products, it becomes imperative to develop sophisticated tools to analyze consumer sentiments and trends effectively.

Aspect-Based Sentiment Analysis (ABSA) plays a crucial role in social listening systems by providing detailed insights into specific aspects of products or services that consumers mention. Unlike traditional sentiment analysis, which provides a general sentiment score for an entire text, ABSA

[1] https://metric.vn/insights/category/metric/ecom-market-research/

breaks down the text into various aspects and assesses the sentiment associated with each aspect. For example, in the beauty industry, a customer review might state: "Màu son đp nhng cht son khô quá" (*The lipstick color is gorgeous, but the texture is too dry*). ABSA would identify two aspects (color and texture) with their respective sentiments (positive for color, negative for texture). This granular level of analysis enables businesses to pinpoint exact areas of strength and weakness in their offerings, leading to more targeted improvements and marketing strategies.

Therefore, in this work, we propose a comprehensive social listening system tailored specifically for beauty products. Our proposed system is designed to leverage data from both text and video sources, incorporating advanced **Normalizer modules** to standardize the information — an effort that marks the first work of its kind. This system performs ABSA to gauge customer sentiment on specific aspects of beauty products, providing granular insights into consumer opinions. Additionally, we attempted to identify mentioned subjects within the data, enabling a more precise understanding of the topics and entities being discussed. We focus on the beauty domain, where product feedback is linguistically diverse and visually driven, making it a challenging and representative case for multimodal social listening.

In light of these developments, we propose a novel social listening system specifically designed for beauty products. Our main contributions can be summarized as follows:

1. We propose a comprehensive social listening system for the beauty industry that integrates both text and video data, with specialized **Normalizer modules** for each data type. To the best of our knowledge, this is the first work to integrate such comprehensive data normalization in a social listening system.

2. We implement an advanced **ABSA module** tailored for the beauty industry, capable of identifying fine-grained aspects and their associated sentiments. This allows for nuanced understanding of consumer opinions on specific product attributes.

3. We pioneer a subject identification feature, aiming to track not only general sentiments but also specific brands being discussed. While the current outputs require further refinement, this represents the first step towards enhancing the granularity of social listening capabilities in the beauty industry.

The rest of this paper is structured as follows: Section 2 explores related work, providing an overview of existing social listening systems and ABSA applications in e-commerce and product reviews. Section 3 outlines our methodology, detailing the design, implementation, data collection, normalization, and analysis processes of our proposed system. Section 4 presents the experimental setup and results, comparing ABSA performance against previous work. Section 5 discusses the limitations of our study, and Section 6 concludes our work.

2 Related Work

Social listening has emerged as a crucial tool for businesses to gain insights from unstructured data on digital platforms. Several studies have demonstrated its effectiveness in various domains. For instance, social listening provides valuable insights into stakeholder perceptions of a company's performance across different dimensions (Westermann and Forthmann, 2021).

Aspect-Based Sentiment Analysis (ABSA) has been widely studied in the context of e-commerce and product reviews. Nasim and Haider (Nasim and Haider, 2017) proposed an ABSA Toolkit for performing aspect-level sentiment analysis on customer reviews. Li et al. (Li et al., 2023) employed aspect-based sentiment analysis of customer-generated content to enhance the prediction of restaurant survival. For Vietnamese, Luc et al. (Luc Phan et al., 2021) built a social listening system based on ABSA for mobile e-commerce.

Recent studies have also addressed normalization in various NLP contexts. For instance, Nguyen et al. (Nguyen et al., 2024) introduced the ViLexNorm corpus for Vietnamese lexical normalization on social media, providing a strong benchmark for transforming informal user-generated text into canonical forms. In parallel, Liao et al. (Liao et al., 2023) investigated post-processing for readability in Automatic Speech Recognition (ASR) transcripts, formulating the task as a sequence-to-sequence generation problem to enhance grammaticality and fluency of spoken-text output. However, to the best of our knowledge, no prior work has integrated both lexical and ASR-based normalization within a unified Vietnamese social listening pipeline.

Our work builds upon these foundations, integrating ABSA, text normalization, and subject identification into a comprehensive social listening system specifically designed for the Vietnamese beauty product market. Text normalization is a critical step in processing social media data and ASR output. Subject identification, particularly brand detection, is an important aspect of social media analytics. To the best of our knowledge, we are the first to develop specialized normalization techniques for both social media comments and ASR output, and attempt to detect the mentioned brands from the beauty product reviews in Vietnamese.

3 Methodology

3.1 Pipeline

Our social listening system for beauty products consists of several interconnected modules designed to process and analyze data from both text and video sources. The overall system architecture is illustrated in Figure 1.

The data flow through the system can be described as follows:

1. **Input**: The system accepts two types of input - Comments Data (text) and Videos Data (audio).

2. **Normalization**:

   - Normalization in our system covers lexical normalization (for slang, abbreviations, and misspellings) and transcript normalization (for ASR disfluencies and missing punctuation). Brand names are also standardized to canonical forms where applicable. For text data, the Comment Normalizer processes the input.
   - For audio data, the audio is first passed through the Automatic Speech Recognition (ASR) module to generate a transcript, which is then processed by the Transcript Normalizer.

3. **ABSA**: The normalized text is fed into the Aspect-Based Sentiment Analysis module, which outputs aspect-polarity pairs for each input.

4. **Subject Identification**: Additionally, the normalized text from both sources is passed through the Subject Identifier module, which extracts mentioned brands.

In our system, Apache Kafka (Kreps et al., 2011) serves as a distributed streaming platform, ingesting two types of input from various sources. Kafka allows us to decouple data producers (e.g., social media platforms, video platforms) from our processing pipeline, ensuring high throughput and fault tolerance. PySpark, the Python API for Apache Spark (Zaharia et al., 2016), is utilized to process the data in a distributed manner. PySpark enables us to perform batch and stream processing on the ingested data, allowing for efficient execution of our pipeline modules across a cluster of machines.

3.2 Modules

Our pipeline consists of five key modules:

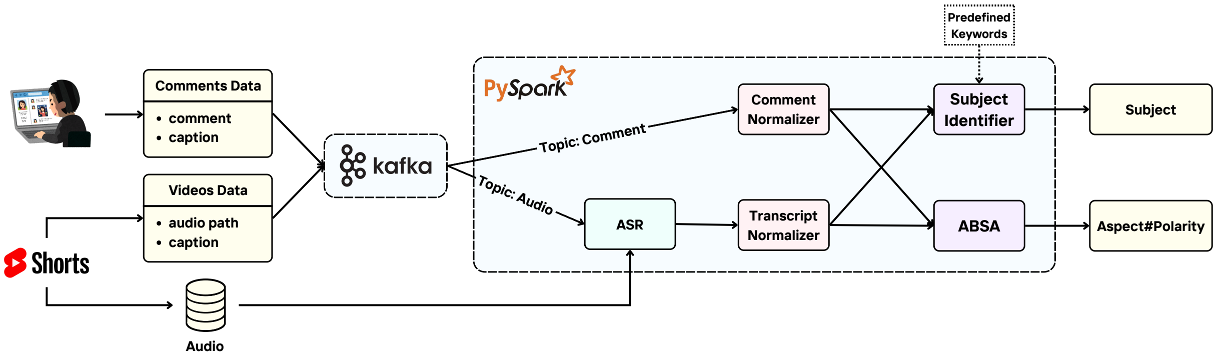
1. **ASR**: We utilize the wav2vec2-base-vietnamese-250h model (Nguyen, 2021) for converting audio to text. This model was pretrained on 13,000 hours of unlabeled Vietnamese YouTube audio and fine-tuned on 250 hours of labeled data from the VLSP 2020 ASR dataset[2]. The ASR challenge associated with this dataset is part of the annual workshop conducted by the Vietnamese Language and Speech Processing (VLSP) community, a prominent event in the field. The transcripts of this dataset were annotated exactly as the audio, retaining all disfluencies and phonetic variations while lacking any punctuations. Therefore, they necessitate further normalization in subsequent steps.

2. **Comment Normalizer**: We employ the BARTPho$_{\text{syllable}}$ base[3] version of BARTpho (Tran et al., 2022a), a pre-trained Sequence-to-Sequence model tailored for Vietnamese to normalize user comments. We trained this model on the ViLexNorm dataset (Nguyen et al., 2024), the first corpus developed for the Vietnamese lexical normalization task. The corpus comprises over 10,000 pairs of sentences meticulously annotated by human annotators, sourced from public comments on Vietnam's most popular social media platforms. This approach ensures robust normalization of

[2] https://vlsp.org.vn/vlsp2020/eval/asr

[3] https://huggingface.co/vinai/bartpho-syllable-base

Figure 1: Our overall social listening system.



colloquial and informal Vietnamese text commonly found in social media comments. For example, the comment "son mac chinh hang nhee" is normalized to "son MAC chính hãng nhé", unifying both orthography and brand capitalization.

3. **Transcript Normalizer**: Similar to the Comment Normalizer, we used the $\text{BARTPho}_{\text{syllable}}$ base model. We created augmented data to pretrain the model then fine-tuned on human-annotated data to enhance its performance. Details about the data can be found in Appendix A. This two-stage process allows the model to handle the unique challenges of normalizing ASR output, including lack of punctuation and potential transcription errors.

4. **ABSA**: Our ABSA module is based on the PhoBERT-base-v2 version of PhoBERT (Nguyen and Tuan Nguyen, 2020). We normalized an ABSE lipstick dataset (Tran et al., 2022b), then finetune the model on it. This dataset contains 16,227 reviews about lipsticks, encompassing a total of 32,775 aspect-sentiment pairs. By focusing on specific aspects of beauty products, such as packaging, texture, and effectiveness, this fine-tuning process enables our model to deliver precise and granular insights into consumer opinions and sentiments. The detailed architecture of the ABSA module is discussed in Section 3.3.

5. **Subject Identifier**: This module enhances the granularity of our social listening system by tracking specific brands mentioned in user comments and reviews. We use a list of predefined keywords and PhoNLP (Nguyen and Nguyen, 2021), a BERT-based multi-task learning model for named entity recognition, to detect brands accurately. Despite the early stage of development and the need for further improvement in output accuracy, this feature represents a significant advancement in social listening capabilities. PhoNLP was selected over general-purpose Vietnamese NER toolkits such as Stanza due to its stronger Vietnamese-specific pretraining and higher accuracy on local benchmarks.

3.3 ABSA Architecture

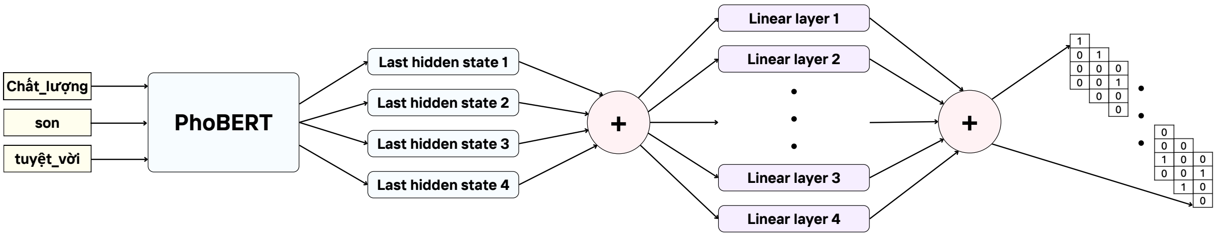
The ABSA module is a crucial component of our system, designed to identify aspects of beauty products and their associated sentiments. Our ABSA architecture is inspired by the multi-task approach for the hotel domain in the paper (Dang et al., 2022). This approach leverages PhoBERT as a pre-trained language model and processes both Aspect Category Detection and Sentiment Polarity Classification tasks simultaneously. We concatenate the last four hidden states of PhoBERT to form a comprehensive representation, which is then fed into separate layers for each task. The ABSA architecture is illustrated in Figure 2.

The architecture is as follows:

- Input Processing: The normalized text is first segmented using the VnCoreNLP toolkit (Vu et al., 2018), which ensures accurate word segmentation for Vietnamese. The segmented text is then tokenized and fed into the PhoBERT model.

- PhoBERT Encoding: The PhoBERT model processes the input and generates contextualized embeddings for each token.

Figure 2: The architecture of our ABSA module.



- Linear Layers: The last hidden states from PhoBERT are passed through a series of linear layers. Each linear layer corresponds to a specific aspect (e.g., colour, texture, smell, price). In our dataset, we define eight fixed aspects: Colour, Texture, Smell, Price, Packing, Staying Power, Shipping, and Others. Each linear head predicts the sentiment of one aspect. For unseen or ambiguous cases, predictions fall into the "OTHERS" category. This fixed-aspect design allows direct comparison with the benchmark lipstick dataset and avoids label drift when training on domain-specific data.

- Output Generation: The output of each linear layer is aggregated to produce the final predictions. For each aspect, the system determines whether it is present in the input and, if so, what the associated sentiment is (positive, negative, or neutral).

This architecture allows for simultaneous detection of multiple aspects and their sentiments within a single input, providing a comprehensive analysis of user opinions on beauty products.

4 Experiment

4.1 Experimental Setup

Each module in our pipeline was trained with specific configurations to ensure optimal performance. The training configurations for all modules are as follows:

- **ASR Module:** We utilized the pre-trained wav2vec2-base-vietnamese-250h model without further fine-tuning for our specific task.

- **Comment Normalizer:**
  - Model: BARTPho$_{\text{syllable}}$ base
  - Learning rate: 5e-05
  - Train batch size: 8
  - Evaluation batch size: 8
  - Random seed: 42
  - Optimizer: Adam
  - Number of epochs: 10

- **Transcript Normalizer:**
  - Model: BARTPho$_{\text{syllable}}$ base
  - Learning rate: 5e-05
  - Train batch size: 8
  - Evaluation batch size: 8
  - Random seed: 42
  - Optimizer: Adam
  - Training strategy: 3 epochs pretraining, followed by 5 epochs fine-tuning

- **ABSA Module:**
  - Model: PhoBERT-base-v2
  - Batch size: 32
  - Learning rate: 2e-5
  - Number of epochs: 10

- **Subject Identifier:** We utilized the PhoNLP model without further fine-tuning for our specific task.

It is important to note that our evaluation focuses specifically on the performance of the ABSA module, as it represents the final output of our pipeline. The evaluation and test results for the ABSA module are presented in Section 4.2.

To evaluate the performance of the ABSA module, we employed the F1 score metric. Weighted averages of F1-score are used to assess the overall performance of the model.

4.2 ABSA Module Results

The experimental results for the ABSA module on the development and test sets of the ABSA lipstick dataset are presented in Table 1. The results from the dataset paper (Tran et al., 2022b) are included

Table 1: ABSA module results on development and test sets (%).

Dataset	$F1_{Aspect}$	$F1_{Sentiment}$
Dev Set	98.46	97.16
Test Set	98.44	97.17
Baseline	*97.51*	*86.92*

for comparison. Note that these scores reflect the best-reported performance for the multi-task learning approach in the literature for the same dataset and serve as the baseline for our experiment.

The results demonstrate the effectiveness of the proposed system. On both the development and test sets, our ABSA module achieves high F1 scores for aspect extraction and sentiment classification, significantly surpassing the results reported in the dataset paper (Tran et al., 2022b). Specifically, our model attains an $F1_{Aspect}$ score of 98.46% on the development set and 98.44% on the test set, compared to 97.51% reported in the dataset paper. Similarly, for $F1_{Sentiment}$, our model achieves scores of 97.16% on the development set and 97.17% on the test set, outperforming the 86.92% reported in the dataset paper.

Our module's performance indicates significant improvements over the baseline, particularly in sentiment classification, where we observe a substantial increase of over 10 percentage points in F1. This improvement can be attributed to the advanced techniques employed in our system: the **Normalizer** modules. We note that both data normalization and the use of augmented samples contribute to improved consistency and reduced noise in the training data. While no separate ablation was conducted, manual inspection suggests normalization plays the dominant role in enhancing sentiment classification accuracy.

Although no formal statistical significance test was conducted, the observed improvements consistently exceeded run-to-run variance, suggesting that the gains are statistically meaningful rather than due to random fluctuations.

In addition to evaluating overall aspect and sentiment detection performance, we also analyzed the performance of our ABSA module for individual aspects. The F1-scores for aspect detection, along with the sample counts for each aspect, are presented in Table 2. These results provide a more detailed view of how well our module performs on specific aspects compared to the baseline.

The table highlights that our ABSA module performs well in most aspects, especially for "COLOUR" and "OTHERS," where it achieved the highest F1-scores of 98.95% and 97.72%, respectively. These aspects also have the highest number of samples, suggesting that the model benefits from having more data for training and evaluation. In addition, the lower performance in "SMELL" and "SHIPPING" can be attributed to the complexity and variability of these specific aspects in the dataset.

Overall, our ABSA module achieves significant advancements in aspect extraction and sentiment classification, surpassing the best-reported benchmarks on the ABSA lipstick dataset (Tran et al., 2022b). However, while excelling in most aspects with ample data, challenges persist in certain aspects, which require further refinement in handling linguistic variability and dataset balance.

5 Limitations

While our aspect-based sentiment analysis module demonstrates strong performance, the **Subject Identifier** component currently offers preliminary results and remains an area for further improvement. Without a comprehensive predefined brand list, the PhoNLP model faces challenges in recognizing newly emerging or less frequent brand names. This limitation reflects the inherent dynamics of the beauty market rather than a constraint of the model itself. To address this, future work will focus on expanding the brand lexicon and exploring adaptive named-entity recognition strategies capable of identifying unseen brands in real time.

Moreover, the current version does not yet associate sentiment polarity directly with identified brands. Integrating brand-level sentiment analysis within the **Subject Identifier** module is a promising next step toward delivering more fine-grained insights into consumer attitudes and market trends.

Finally, although the evaluation has been conducted primarily on beauty product reviews, the proposed architecture is domain-agnostic and can be readily extended to other sectors (e.g., fashion, electronics) with minimal retraining. This generalizable design highlights the scalability of our pipeline beyond the current experimental domain.

6 Conclusion

In this study, we have proposed and implemented a specialized social listening system tailored for the

Table 2: The F1-score (%) for aspect detection in each aspect with sample count

Aspect	Baseline	Our Module	Sample Count
SMELL	98.01	97.00	363
COLOUR	96.60	**98.95**	9720
STAYINGPOWER	95.67	**95.90**	464
PRICE	96.37	**96.88**	324
SHIPPING	97.15	96.10	530
PACKING	95.89	**96.72**	316
TEXTURE	94.88	**95.26**	460
OTHERS	94.68	**97.72**	754

beauty industry, leveraging advanced techniques in Aspect-Based Sentiment Analysis (ABSA), data normalization, and subject identification. Our system integrates textual and video data sources, aiming to extract detailed insights into consumer sentiments and trends surrounding beauty products. Through our experimental evaluations, we have demonstrated the effectiveness of our approach. The ABSA module achieved significant advancements in aspect extraction and sentiment classification compared to existing work. Furthermore, the incorporation of data normalization modules tailored for Vietnamese text and video transcripts has proven crucial in enhancing the accuracy and reliability of sentiment analysis outputs. We also highlights the importance of subject identification in social listening systems, particularly in tracking mentions of brands and specific entities within consumer discussions. While our current model shows promising results, further research and refinement are needed to adapt to evolving brand landscapes and improve accuracy in real-time scenarios. We hope that this research contributes a comprehensive framework for applying social listening and ABSA methodologies to gain deeper insights into consumer behaviors and preferences within the dynamic beauty market. Future work will focus on extending the system to additional domains and performing significance testing to further verify the robustness of the observed improvements.

Acknowledgments

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

References

Hoang-Quan Dang, Duc-Duy-Anh Nguyen, and Trong-Hop Do. 2022. Multi-task solution for aspect category sentiment analysis on vietnamese datasets. In *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pages 404–409.

Jay Kreps, Neha Narkhede, Jun Rao, and 1 others. 2011. Kafka: A distributed messaging system for log processing. In *Proceedings of the NetDB*, volume 11, pages 1–7. Athens, Greece.

Hengyun Li, XB Bruce, Gang Li, and Huicai Gao. 2023. Restaurant survival prediction using customer-generated content: An aspect-based sentiment analysis of online reviews. *Tourism Management*, 96:104707.

Junwei Liao, Sefik Eskimez, Liyang Lu, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng. 2023. Improving readability for automatic speech recognition transcription. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(5).

Luong Luc Phan, Phuc Huynh Pham, Kim Thi-Thanh Nguyen, Sieu Khai Huynh, Tham Thi Nguyen, Luan Thanh Nguyen, Tin Van Huynh, and Kiet Van Nguyen. 2021. Sa2sl: From aspect-based sentiment analysis to social listening system for business intelligence. In *Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Tokyo, Japan, August 14–16, 2021, Proceedings, Part II 14*, pages 647–658. Springer.

Wendy W Moe and David A Schweidel. 2017. Opportunities for innovation in social media analytics. *Journal of product innovation management*, 34(5):697–702.

Zarmeen Nasim and Sajjad Haider. 2017. Absa toolkit: An open source tool for aspect based sentiment analysis. *International Journal on Artificial Intelligence Tools*, 26(06):1750023.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online. Association for Computational Linguistics.

Linh The Nguyen and Dat Quoc Nguyen. 2021. PhoNLP: A joint multi-task learning model for Vietnamese part-of-speech tagging, named entity recognition and dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–7.

Thai Binh Nguyen. 2021. Vietnamese end-to-end speech recognition using wav2vec 2.0.

Thanh-Nhi Nguyen, Thanh-Phong Le, and Kiet Nguyen. 2024. ViLexNorm: A lexical normalization corpus for Vietnamese social media text. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1421–1437, St. Julian's, Malta. Association for Computational Linguistics.

Nguyen Luong Tran, Duong Minh Le, and Dat Quoc Nguyen. 2022a. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association.*

Quang-Linh Tran, Phan Thanh Dat Le, and Trong-Hop Do. 2022b. Aspect-based sentiment analysis for Vietnamese reviews about beauty product on E-commerce websites. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 767–776, Manila, Philippines. Association for Computational Linguistics.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

Arne Westermann and Jörg Forthmann. 2021. Social listening: a potential game changer in reputation management how big data analysis can contribute to understanding stakeholders' views on organisations. *Corporate Communications: An International Journal*, 26(1):2–22.

Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. 2016. Apache spark: a unified engine for big data processing. *Commun. ACM*, 59(11):56–65.

A Details of Data Used for Transcript Normalization

A.1 Human-annotated Data for Finetuning

To construct our gold dataset, we leveraged a 100-hour speech public dataset from Vinbigdata[4], which serves as a clean subset of the VLSP2020 ASR competition[5]. Specifically, this dataset comprises 100 hours of speech data collected from open sources and manually transcribed with an impressive accuracy rate of 96%.

The Vinbigdata dataset comprises a total of 56,427 samples. We excluded samples with fewer than four characters to maintain quality. The transcripts of this dataset were annotated exactly as the audio, retaining all disfluencies and phonetic variations while lacking any punctuations. This raw, unprocessed data formed the basis of our gold dataset.

To generate readable target transcripts, we normalized the original transcripts from the dataset of Vinbigdata. This normalization involved removing deletable portions of disfluencies and fillers, shortening abbreviations, and normalizing phonetic variations to correspond to common Vietnamese words. Specifically, we annotated 596 pairs for training, 250 for the development set, and 250 for the test set. Importantly, all four types of post-processing (repetitions, fillers, abbreviation, and phonetic variations) were evenly distributed across these sets. The remaining samples were reserved for data augmentation purposes.

This process ensured that our gold dataset not only reflected the original transcription characteristics but also captured the nuances of ASR model errors.

A.2 Augmented Data for Pretraining

Given the time and effort-intensive nature of manual annotation, we explored data augmentation strategies for ASR post-processing task. Our approach involves two key techniques aimed at enhancing the diversity and robustness of our training data. We pretrained the Transcript Normalizer using a combined 30,000 augmented samples: 15,000 generated through predictions and 15,000 synthesized.

A.2.1 Utilizing ASR Model Predictions:

We leveraged the predictions a the ASR model as an augmentation strategy, specifically focusing on the incorrect predictions. To identify errors in the remaining data, we employed a state-of-the-art Vietnamese ASR model[6]. This model, fine-tuned

[4] https://institute.vinbigdata.org/events/vinbigdata-chia-se-100-gio-du-lieu-tieng-noi-cho-cong-dong/

[5] https://vlsp.org.vn/vlsp2020/eval/asr

[6] https://huggingface.co/khanhld/wav2vec2-base-vietnamese-160h

on approximately 160 hours of diverse Vietnamese speech data, has not yet incorporated a language model but has yielded promising results. Out of the remaining 48,927 samples, we randomly selected 15,000 samples where the model predictions were incorrect, thus capturing the errors inherent in the ASR system. We paired the incorrectly predicted samples with their corresponding transcripts from the Vinbigdata dataset. By doing so, we introduced more varied forms of mistakes that an ASR system may generate, enriching our training data and improving the model's ability to handle diverse errors.

A.2.2 Synthetic Data Generation:

In addition to using model predictions, we employed a synthetic data generation approach. We collected the most popular articles from Vietnamese Wikipedia[7] in the year 2023. After gathering the text, we pre-processed it by removing all characters except those in the alphabet or numerical characters. The entire text was then converted to lowercase, and random segments of text, each with a length of fewer than 130 tokens, were extracted to simulate the punctuation-free transcripts in the Vinbigdata dataset. To maintain quality, segments with fewer than four characters were excluded. Afterwards, we randomly selected 15,000 samples.

Subsequently, we performed synthetic data augmentation by simulating repetitions, fillers, and phonetic variations. To prevent confusion, only one of the three synthetic methods was applied to each sample, selected randomly with equal probabilities (33.33%). The synthesis methods were implemented as follows:

1. **Repetitions:**
   - Examine each token and decide whether to introduce a repetition with a 30% probability.
   - Special case: If the token being assessed is a conjunction or connective, the decision probability increases to 50%.
   - Implement repetitions: 80% of instances involve duplication, 15% triplication, and 5% quadruplication.
2. **Fillers:**
   - Examine each token and determine whether to include a filler with a 20% probability.
   - Implement: Add one token after the chosen token with the following probabilities:
     - 70% for selecting from frequently used filler words.
     - 30% for selecting from less common filler words.
3. **Phonetic Variations:**
   - Examine each token and decide whether to apply a phonetic variation with a 30% probability.
   - Implement:
     - Analyse the first and last characters of the token to identify whether they belong to a list of words prone to phonetic confusion.
     - If affirmative, replace one of the two, either the beginning or the end of the token, with a probability of 50%.

Additionally, in (1) and (2), we employed word segmentation using VnCoreNLP (Vu et al., 2018) (applied to 50% of samples) before iterating through tokens. This introduced cases where repetitions span both simple and compound words in (1) and fillers may or may not be inserted between compound words in (2).

[7] https://vi.wikipedia.org/