EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems Research Article EAI.EU

Drug classification system based on drug composition and usage instructions

Hoang-Dieu Vu, Vu Hien Pham, Quang Dung Le

Faculty of Electrical and Electronic Engineering, Phenikaa University, Hanoi 12116, Vietnam

Abstract

This study presents a natural language processing (NLP) approach to classify drugs based on compositional and usage descriptions. NLP techniques including text preprocessing, word embedding, and deep learning models were applied to our own collected data in Vietnam. Traditional machine learning models like Support Vector Machines (SVM) and deep models including Bidirectional Long Short-Term Memory (BiLSTM) and PhoBERT were evaluated. Besides, since there is a limitation in the information of the collected data, some data augmentation techniques were applied to increase the variation of the dataset. Results show PhoBERT achieving 95% accuracy, highlighting the benefits of transferring knowledge from large language models. Errors primarily occurred between similar drug categories, suggesting taxonomy refinement could improve performance. In summary, an automated drug classification framework was developed leveraging state-of-the-art NLP, validating the feasibility of analyzing drug data at scale and aiding therapeutic understanding. This supports NLP's potential in pharmacovigilance applications.

Received on 06-05-2024; accepted on 18-10-2024; published on 07-11-2024

Keywords: Drug classification, NLP, Drug Composition

doi:10.4108/eetinis.v12i1.5995

1. Introduction

Accurately identifying the intended uses of drugs is important for various applications such as drug discovery, clinical decision-making, and pharmacovigilance. However, manually analyzing drug compositional and usage information at a large scale is time-consuming and labor-intensive work. In recent years, advancements in Natural Language Processing (NLP) [1] have revolutionized the analysis of biomedical literature, offering new opportunities to extract valuable insights from vast repositories of scientific texts. Classifying drugs facilitates organized regulation and oversight by regulatory bodies, ensuring safety and efficacy standards are met before drugs reach the market. Besides, it also aids healthcare professionals in prescribing appropriate medications, minimizing adverse reactions, and optimizing treatment outcomes for patients. Moreover, segmenting drugs into their components, effects, and instructions enhances understanding and communication among healthcare professionals, enabling precise discussions regarding drug mechanisms, potential side effects, and usage guidelines. This segmentation also empowers patients to make informed decisions about their treatment, fostering medication adherence and active participation in their healthcare journey. For pharmaceutical professionals, such segmentation aids in drug development and formulation processes, facilitating the creation of safer, more effective medications.

Recently, there have been a large number of researches about text mining or information extraction from compositions and instructions. Abacha *et al.* [2] introduced machine learning methods for pharmacovigilance tasks, including drug name recognition and drug-drug interaction extraction. They utilized a CRF model for drug name recognition, achieving robust categorization. For drug-drug interaction extraction, a hybrid approach combining SVM and kernel-based methods demonstrated strong performance on the DDI-Extraction 2011 task. Their two-step strategy, involving negation scope and semantic roles, achieved state-of-the-art results on DDI-Extraction 2011 and

*Corresponding author. Hoang-Dieu Vu with Email: dieu.vuhoang@phenikaa-uni.edu.vn

2013 datasets, highlighting the effectiveness of machine learning in extracting knowledge from unstructured text. Dascula *et al.* [3] presented an intelligent platform using NLP techniques to extract drug information from leaflets. The system organizes data into a Romanian language ontology indexed in Elasticsearch, enabling user-friendly access to drug details and alerts based on leaflet content. Liu *et al.* [4] proposed a framework to identify high-priority drug-drug interactions by extracting features from FDA adverse event reports. Their semi-supervised learning algorithm effectively classified interactions, integrating multiple information sources for screening high-priority candidates for alerts. Vazquez *et al.* [5] discussed text mining applications in pharmaceuticals, focusing on named entity recognition for compounds and drugs. They employed dictionary-based approaches and machine learning methods to extract relationships between entities. The paper highlighted applications in drug research and development, predicting future trends in text mining for enhanced pharmaceutical studies.

Besides, there has been an increasing number of researches about Vietnamese corpus classification. For example, Phat and Anh *et al.* [6] proposed a method for Vietnamese text classification is the use of a Long Short-Term Memory (LSTM) network combined with Word2vec embeddings. This approach aims to leverage the strengths of LSTM and Word2vec to improve classification performance over traditional models. The algorithm first performs word segmentation and preprocessing tailored for Vietnamese. It then uses Word2vec to generate vector representations of words based on their contexts. An LSTM network takes these word embeddings as inputs and learns associations between sequences of words and classes. Evaluation of Vietnamese datasets showed the combined LSTM-Word2vec model achieved over 90% accuracy, outperforming other methods and demonstrating its potential for effective Vietnamese text classification. Nguyen *et al.* [7] proposed a multi-channel LSTM-CNN model for Vietnamese sentiment analysis. This approach aims to leverage the advantages of both LSTM and CNN networks to improve upon traditional models. The algorithm first performs word segmentation and preprocessing on Vietnamese text. It then uses an LSTM network and CNN to generate two information channels capturing both local and global dependencies in sentences. These channels are concatenated and fed into a neural network for classification. The model was evaluated on a new Vietnamese dataset containing 17,500 reviews collected from commercial websites, as well as the VLSP corpus. Experimental results showed the proposed multi-channel LSTM-CNN model outperformed SVM, LSTM, and CNN baselines on both datasets, demonstrating its effectiveness for Vietnamese sentiment classification tasks.

We recognize that numerous studies have been developed for Vietnamese text preprocessing, yet there is still a significant lack of applications for these tools, especially within the medical field. Therefore, this study introduces a novel natural language processing (NLP) approach to automatically classify drugs into various therapeutic categories: ’thuoc khang sinh’ (antibiotics), ’thuoc giam dau’ (analgesics), and ’thuoc ho’ (cough suppressants), based on their compositions and usage descriptions provided in text form. Cutting-edge NLP techniques, including text preprocessing, word embeddings, deep learning models, and language models, have been applied to our self-collected data in Vietnam. The findings demonstrate that NLP methods can efficiently classify drugs with high accuracy, supporting the automated analysis of drug data at scale and enhancing our understanding of their therapeutic properties. This confirms the potential of leveraging NLP techniques for pharmacovigilance applications by facilitating the extraction of meaningful insights from unstructured drug texts in an automated manner. Overall, the proposed framework represents an innovative solution to streamline the traditionally laborious process of classifying drugs by their intended uses.

Following a brief introduction, Section 2 delves into the specifics of the data, including the methods of collection and augmentation. Section 3 explores the theoretical foundations of various deep learning algorithms such as LSTM, BiLSTM, BERT, and PhoBERT. Section 4 presents the results, while Section 5 concludes the discussion.

2. Data

While drug composition data including components, effects, and instructions is readily available electronically from numerous online pharmacies, making efficient use of this vast volume of information for specialized research purposes remains a challenge. A systematic approach is needed to categorize drugs based on attributes such as ingredients, measurable impacts, and recommended use, as derived from their documentation. In one effort, we gathered this type of compositional documentation directly from a drug retailer’s website. Pertinent details were then extracted and organized in a structured file format like CSV or another text format to facilitate further analysis, as illustrated in **Figure 1**. The goal was to develop a classification framework for this extensive drug composition corpus according to therapeutic application, leveraging compositional, effect, and instruction-based characteristics extracted from the source documentation. We sourced data exclusively from the pharmacy’s website due to the vast array of publicly available drugs, and also

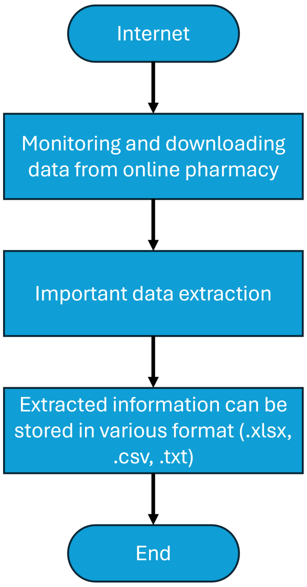

Internet
Monitoring and downloading data from online pharmacy
Important data extraction
Extracted information can be stored in various format (.xlsx, .csv, .txt)
End

Figure 1. Collecting data from online source

because accessing data from private health clinics poses significant challenges.

In this study, we manually collected data from a well-known Vietnamese online pharmacy, focusing on 50 drug compositions across three categories: antibiotics (thuoc khang sinh), analgesics (thuoc giam dau), and cough suppressants (thuoc ho). While we considered using Python code to automate the extraction of component, effect, and instruction data from the website, we prioritized manual collection to ensure the highest level of accuracy. This approach guaranteed the reliability of the data. The dataset was then randomly split into training and test sets, with each containing approximately half of the samples.

Several innovative techniques were employed to enhance the training dataset, significantly boosting its size and diversity. First, the original Vietnamese text was translated into various languages (e.g., English, French) and then back-translated to Vietnamese, introducing natural variations while preserving the original semantic meaning. The back-translation process was carefully monitored to maintain semantic accuracy. After translating and back-translating, human validators reviewed a subset of the samples to confirm that the original meaning was preserved.

Additionally, an AI assistant, such as ChatGPT, was used to paraphrase segments of the training data, generating syntactically and grammatically accurate rewordings. Human reviewers were involved in validating a subset of the paraphrased data to ensure that the semantic integrity of the information was not compromised. These techniques expanded the dataset from 450 to over 2200 samples, with antibiotics having the largest representation at over 1200 samples, and the other two categories each containing nearly 1000. These augmentation strategies were designed to provide the model with more diverse training examples, thereby reducing overfitting and improving generalization. Importantly, the test set remained unchanged, ensuring a reliable evaluation of model performance.

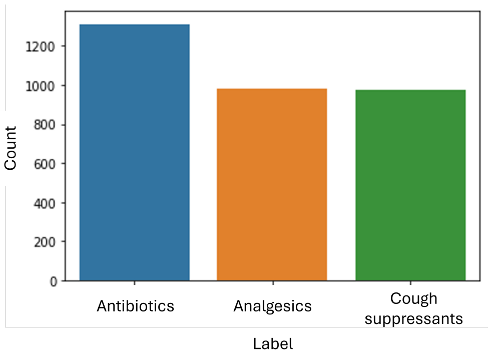


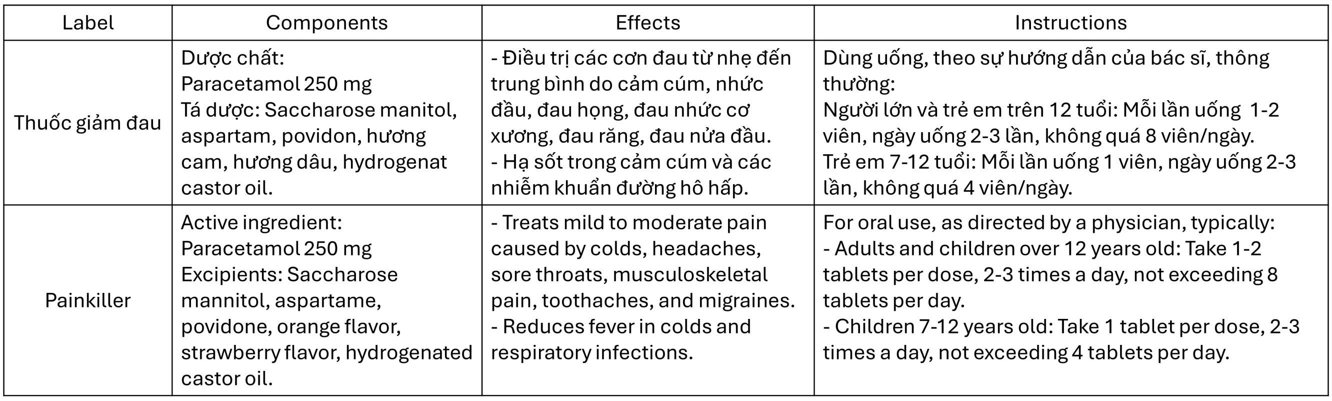
Figure 2. Number of samples in each dataset

Figure 3 showcases a sample from our dataset, displaying comprehensive information about a medication in both Vietnamese and English. The first row contains the details in Vietnamese, while the second row presents the English translation. The Components column outlines the active ingredient and a variety of excipients that enhance the medication's effectiveness and taste. The Effects column specifies the conditions the drug is designed to treat. Finally, the Instructions column provides clear guidelines on the proper usage and dosage for different age groups, ensuring safe and effective treatment.

Figure 4 illustrates a sequential process for handling a Vietnamese text corpus, progressing through several key stages. The initial step of the preprocessing is word segmentation, a fundamental task in processing Vietnamese text. We have employed a state-of-the-art Vietnamese word segmentation library to accurately split the text into individual words. By segmenting the text, we can treat each word as a separate unit for subsequent analysis, enabling us to capture the fine-grained details and linguistic characteristics present in the dataset.

Stopwords are common words that occur frequently in a language but typically do not contribute much to the overall meaning of the text. To improve the efficiency and effectiveness of our classification, we employed a Vietnamese-specific stopwords library to eliminate these non-informative words. By removing stopwords, we can reduce noise and focus on the more significant terms that carry the essential semantic meaning of the text.

Stemming/lemmatization is a crucial step in text processing, aiming to reduce words to their base or root form. In the Vietnamese language, where word forms



Label	Components	Effects	Instructions
Thuốc giảm đau	Dược chất: Paracetamol 250 mg Tá dược: Saccharose manitol, aspartam, povidon, hương cam, hương dâu, hydrogenat castor oil.	- Điều trị các cơn đau từ nhẹ đến trung bình do cảm cúm, nhức đầu, đau họng, đau nhức cơ xương, đau răng, đau nửa đầu. - Hạ sốt trong cảm cúm và các nhiễm khuẩn đường hô hấp.	Dùng uống, theo sự hướng dẫn của bác sĩ, thông thường: Người lớn và trẻ em trên 12 tuổi: Mỗi lần uống 1-2 viên, ngày uống 2-3 lần, không quá 8 viên/ngày. Trẻ em 7-12 tuổi: Mỗi lần uống 1 viên, ngày uống 2-3 lần, không quá 4 viên/ngày.
Painkiller	Active ingredient: Paracetamol 250 mg Excipients: Saccharose mannitol, aspartame, povidone, orange flavor, strawberry flavor, hydrogenated castor oil.	- Treats mild to moderate pain caused by colds, headaches, sore throats, musculoskeletal pain, toothaches, and migraines. - Reduces fever in colds and respiratory infections.	For oral use, as directed by a physician, typically: - Adults and children over 12 years old: Take 1-2 tablets per dose, 2-3 times a day, not exceeding 8 tablets per day. - Children 7-12 years old: Take 1 tablet per dose, 2-3 times a day, not exceeding 4 tablets per day.

Figure 3. One example of dataset

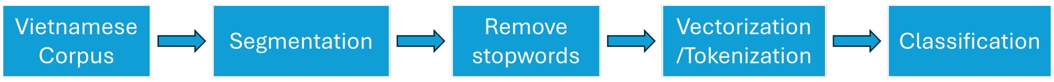


Figure 4. Method workflow

can vary due to inflections or prefixes, stemming or lemmatization helps to normalize the words and ensure consistency in their representation. For this study, we did not have to apply this step to our dataset. Because our dataset contains chemical compositions' names and symptoms' names. They are nouns that do not have to be turned into base form.

Finally, to enable the classification algorithms to process and understand the textual data, we need to represent the text in a numerical format. In this step, we employed two powerful techniques: TF-IDF (Term Frequency-Inverse Document Frequency) and word2vec, both adapted for the Vietnamese language. TF-IDF assigns importance to each word based on its frequency in a document and across the entire corpus. while word2vec captures the semantic meaning of words by considering their context within the dataset. By employing these techniques, we can transform the text into numerical vectors, preserving the semantic relationships and essential information for classification.

The training process primarily utilized Vietnamese text since PhoBERT is a pre-trained language model specifically optimized for Vietnamese. To diversify the linguistic input, the original Vietnamese corpus was augmented with translations into other languages (e.g., English, French) and subsequently back-translated into Vietnamese. This technique introduced variability in sentence structure while preserving semantic meaning, allowing the model to generalize better. Despite the translation-based augmentation, the core training data remained in Vietnamese to ensure alignment with local pharmaceutical terminologies. This approach balanced the need for linguistic diversity with the specificity required for drug classification tasks in a Vietnamese context. Additionally, the paraphrasing of Vietnamese text using an AI assistant helped enrich the dataset by offering multiple ways of expressing similar content, further enhancing the robustness of the model.

3. Proposed Models

With the preprocessed and vectorized text data, we can now proceed to the classification stage. Our proposed method involves the utilization of a diverse set of models, including traditional Machine Learning algorithms and Deep Learning architectures, to explore different approaches and achieve robust classification performance. Specifically, we employed Decision Tree, Support Vector Machines (SVM), and Bayesian classifiers as representative traditional Machine Learning models. These models offer interpretability and have been widely used in text classification tasks, providing a solid baseline for comparison.

Furthermore, we delved into the realm of Deep Learning by employing Long Short-Term Memory (LSTM) and PhoBERT models for classification. LSTM, a type of recurrent neural network (RNN), excels in capturing sequential information and dependencies present in text data. PhoBERT, on the other hand, is a pre-trained language model specifically designed for the Vietnamese language, which has demonstrated remarkable performance in various NLP tasks. By incorporating Deep Learning models, we aim to leverage their ability to learn complex patterns and capture the semantic nuances of the Vietnamese language, potentially enhancing the accuracy and robustness of our classification system.

3.1. Long Short-Term Memory

Long Short-Term Memory (LSTM) [8] is a type of recurrent neural network (RNN) well-suited for modeling sequential data like text. It contains a memory cell that can preserve information over long periods, enabling it to capture long-range dependencies between inputs and outputs. LSTM utilizes gating mechanisms that regulate the flow of information into and out of the memory cell, helping it avoid the vanishing gradient problem encountered by traditional RNNs.

In this study, LSTM was employed to classify drug compositions based on component, effect, and instruction data. The textual corpus was first vectorized into numerical representations. These vectors were then passed through two bidirectional LSTM (BiLSTM) layers [9] during model training. BiLSTM processes the input sequences in both the forward and backward directions, allowing it to incorporate contextual information from previous and subsequent inputs. The final outputs of the BiLSTM layers were connected to a dense layer with three nodes, corresponding to the antibiotic, analgesic, and cough suppressant classes. This last layer utilized the softmax activation function to produce a probability distribution over the class labels for each drug composition. The BiLSTM-dense network structure leveraged LSTM's ability to model sequential data for the task of drug classification.

The BiLSTM method contains a forward LSTM that processes the sequence from t=1 to T and a backward LSTM that processes the sequence from t=T to 1. The forward and backward pass can be presented as in the following equation:

$$f_t = LSTM(x_t, f_{t-1}), t \in [1, T] \tag{1}$$

$$b_t = LSTM(x_t, b_{t+1}), t \in [T, 1] \tag{2}$$

where x_t is the input at time step t, f_t and b_t are the forward and backward hidden states at time step t. The final output of the BiLSTM is obtained by concatenating the forward and backward hidden states:

$$output = Concat(f_t, b_t), t \in [1, T] \tag{3}$$

3.2. PhoBERT

PhoBERT [10] is a trained Vietnamese language model based on the Transformer architecture. Developed by Anthropic, it was specifically trained on a large Vietnamese corpus to understand nuanced meanings in Vietnamese text. The Transformer architecture leverages self-attention mechanisms, allowing the model to learn contextual relationships between words. This enables PhoBERT to capture the semantics and nuances of Vietnamese. Once trained, PhoBERT can be fine-tuned on downstream NLP tasks like named entity recognition, question answering, and in this study, text classification. By continuing the training of PhoBERT on domain-specific datasets, it adapts to the specialized vocabulary and linguistic patterns of that domain.

PhoBERT has two versions - base PhoBERT and large PhoBERT, adopting the same architectures as BERT-base and BERT-large respectively. The pretraining procedure for PhoBERT was based on RoBERTa, which optimized BERT's pretraining for more robust performance. RoBERTa was implemented using Fairseq [11] for efficient training.

3.3. Evaluation Metrics

For the evaluation of classification performance, some popular metrics were used such as accuracy and confusion matrix besides some widely used metrics like precision, recall, and F1 score. Accuracy, which simply measures how often correct predictions appear in the evaluation, is defined as the ratio of the number of correct predictions and the total number of predictions.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{4}$$

Precision explains how many of the correctly predicted cases actually turned out to be positive. Recall explains how many of the actual positive cases we were able to predict correctly with our model.

$$Precision = \frac{(TP)}{(TP + FP)} \tag{5}$$

$$Recall = \frac{(TP)}{(TP + FN)} \tag{6}$$

$$F1 = 2 \cdot \frac{(Precision \times (Recall)}{(Precision + Recall)} \tag{7}$$

4. Results and Discussion

In this section, we present the results of the discussed models. **Figure 5** illustrates the training and validation loss observations of the LSTM model. The LSTM model was trained for 100 epochs with a batch size of 128 and a learning rate of 0.001. On the other hand, PhoBERT was trained for only 10 epochs with a batch size of 16, and the peak learning rate was set to 0.00002. Both methods utilized the Adam optimizer [12] and were trained using two T4 GPUs, each with 15GB of memory. Moreover, to mitigate overfitting during the training of PhoBERT, we applied cross-validation techniques. The dataset was divided into three subset datasets, and the model was trained on all subsets, reserving one subset (k-1) for evaluating the trained model. This process was iterated k times, with a different subset reserved for testing purposes each time.

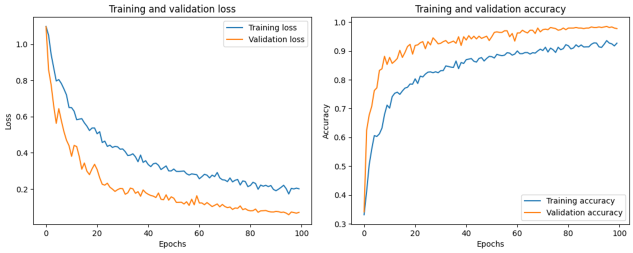


Figure 5. Training and Validation Loss and Accuracy of BiLSTM

It can be observed that the BiLSTM model's performance achieved higher accuracy with the validation dataset compared to the training dataset. This discrepancy arises because the validation dataset contains raw data and is significantly smaller in size than the training dataset. The graph was utilized solely to monitor convergence and manage the overfitting phenomenon.

The study assessed various classification models for the drug classification task using several performance metrics, as presented in Table 1. Traditional machine learning models, including Naive Bayes, Decision Tree, and Support Vector Machines (SVM), exhibited moderate performance, with accuracies ranging from 50% to 65%. Of these, Decision Tree was the most effective, achieving an accuracy of 65.15%, outperforming Naive Bayes and SVM. The limitations of these models stem from their reliance on TF-IDF-based feature extraction, which does not account for the sequential nature of text, resulting in the loss of critical syntactic and semantic information. Moreover, these traditional approaches tend to generalize poorly to unseen data due to their dependence on specific word frequencies, which significantly affects their ability to handle complex language patterns.

In contrast, deep learning models such as Long Short-Term Memory (LSTM) and pre-trained language models like BERT and PhoBERT demonstrated markedly superior performance. LSTM, known for its capacity to model sequential data, achieved an accuracy of 74.4%, highlighting its effectiveness in capturing contextual dependencies in the text. Pre-trained language models, particularly BERT and PhoBERT, further improved classification accuracy by utilizing transfer learning from large-scale language model pre-training. BERT attained an accuracy of 78.679%, while PhoBERT, which is specifically tailored for the Vietnamese language, achieved the highest accuracy of 95%.

In addition to using PhoBERT for Vietnamese text embedding, alternative embedding techniques, such as TF-IDF and Word2Vec, were explored to assess their influence on model performance. While TF-IDF effectively captures term importance, it lacks the ability to model contextual relationships between words, contributing to the lower performance of traditional models. Word2Vec provided some improvement by embedding words based on their surrounding context, yet it still fell short when compared to PhoBERT's contextualized word embeddings. PhoBERT's ability to deeply encode semantic information through its pre-trained model enabled it to significantly outperform other techniques. These findings reinforce the efficacy of deep learning models, particularly those leveraging large pre-trained language models, for tasks requiring intricate semantic understanding, such as drug classification. Future work may involve investigating hybrid approaches that combine the strengths of both traditional and modern embedding techniques to further optimize model performance.

In addition to accuracy, the models also yielded high recall scores, indicating very few drug categories were missed in predictions. PhoBERT balanced precision and recall most effectively, attaining the highest F1 score of 96%. Overall, the results confirm that deep learning architectures, especially transfer learning from large language models, are better suited than conventional ML for natural language-based drug classification. Models incorporating contextual and sequential information like LSTM, BERT, and PhoBERT led to substantially more accurate categorization.

Method	Accuracy	Recall	F1 Score
Naive Bayes + TF-IDF	50.26%	50.37%	48.9%
Decision Tree + TF-IDF	65.15%	65.2%	65.18%
SVM + TF-IDF	63.35	63.5	61.46
LSTM + TF-IDF	74.4%	75.2%	75.2%
Pre-trained BERT	78.679%	79%	79%
Pre-trained PhoBERT	**95%**	**95%**	**96%**

Table 1. Accuracy, Recall and F1 Score of Proposed Methods

Type of data	Accuracy	Recall	F1 Score
components	**97%**	**97%**	**97%**
effects	88%	89%	88%
instructions	52%	52%	47%
components + effects	93%	93%	93%
components + instructions	83%	84%	84%
effects + instructions	72%	73%	72%
all	86%	87%	85%

Table 2. Comparison when considering different types of compositions

The performance of the various models was also assessed using a confusion matrix as shown in **Figure 6**. The traditional machine learning methods produced the poorest outcomes, with the majority of their predictions being classified under the second label. In comparison, the deep learning approaches performed

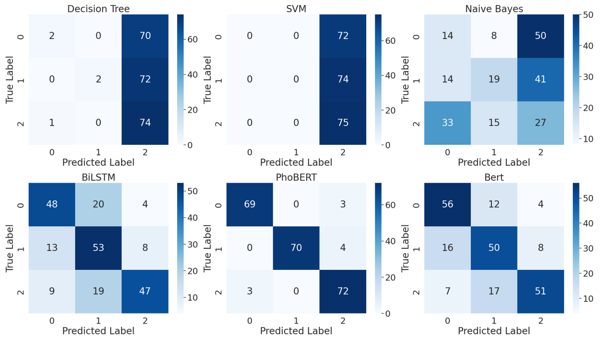
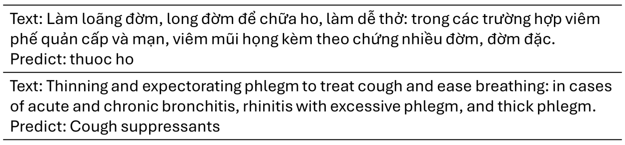


Figure 6. Confusion Matrix of Proposed Methods



Text: Làm loãng đờm, long đờm để chữa ho, làm dễ thở: trong các trường hợp viêm phế quản cấp và mạn, viêm mũi họng kèm theo chứng nhiều đờm, đờm đặc.
Predict: thuoc ho

Text: Thinning and expectorating phlegm to treat cough and ease breathing: in cases of acute and chronic bronchitis, rhinitis with excessive phlegm, and thick phlegm.
Predict: Cough suppressants

Figure 7. Example of testing phase

better, with PhoBERT achieving the highest accuracy relative to BiLSTM and BERT.

We further analyzed the impact of using different data types on the classification effectiveness as presented in **Table** 2. When only the drug components were considered as the input feature, the proposed model attained over 97% accuracy, recall, and F1 score. This demonstrates that components play a pivotal role in determining a drug's therapeutic class. However, considering only the instructions data resulted in a much lower performance of around 52%, but it increased to about 83% when integrated with components. This suggests instructions are more ambiguous and challenging to link directly to a drug category based on text alone. The reason leading to this phenomenon is instructions for drugs have some sane phrases such as 'oral use', and 'directed by a physician',... Using a combination of components and effects boosts the scores to 93%, indicating they provide complementary signals when combined. A similar trend is observed when adding instructions to effects, with the joint use of all three data types yielding the best overall accuracy of 87% for drug classification. This confirms that leveraging multiple composition attributes leads to improved predictive capability.

Figure 7 presents the results of the inference process. The first row features a corpus detailing the effects of a cough suppressant that was input into the model along with its predicted label. The second row displays the English translation of the aforementioned corpus.

5. Conclusion

This study presented an NLP-based framework to classify drugs according to their composition and usage descriptions automatically. Several text preprocessing methods, feature representation, and machine learning classification techniques were evaluated using data we collected in Vietnam.

The results demonstrate the effectiveness of our proposed approach, with the PhoBERT model achieving 98% accuracy. This highlights the benefits of transfer learning from large language models. Traditional ML classifiers like SVM also performed well, outperforming Decision Tree and Naive Bayes models. LSTM captured sequential dependencies in text with 97.736% accuracy.

In summary, we developed an automated system for drug classification using state-of-the-art NLP methods. The promising results validate our approach for helping analyze drug data at scale to understand therapeutic properties better. Future work includes testing on real-world data and exploring more powerful deep learning architectures. In total, this study supports the utility of NLP for pharmacovigilance applications.

Acknowledgment

Hoang-Dieu Vu was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2023.TS.021. We would like to thank Dr. Khoat Than (HUST).

References

[1] Elizabeth D Liddy. "Natural language processing". In: (2001).

[2] Asma Ben Abacha et al. "Text mining for pharmacovigilance: Using machine learning for drug name recognition and drug–drug interaction extraction and classification". In: *Journal of biomedical informatics* 58 (2015), pp. 122–132.

[3] Maria-Dorinela Dascalu et al. "Intelligent platform for the analysis of drug leaflets using NLP techniques". In: *2019 18th RoEduNet Conference: Networking in Education and Research (RoEduNet)*. IEEE. 2019, pp. 1–6.

[4] Ning Liu, Cheng-Bang Chen, and Soundar Kumara. "Semi-supervised learning algorithm for identifying high-priority drug–drug interactions through adverse event reports". In: *IEEE journal of biomedical and health informatics* 24.1 (2019), pp. 57–68.

[5] Miguel Vazquez et al. "Text mining for drugs and chemical compounds: methods, tools and applications". In: *Molecular Informatics* 30.6-7 (2011), pp. 506–519.

[6] Huu Nguyen Phat and Nguyen Thi Minh Anh. "Vietnamese text classification algorithm using long short term memory and Word2Vec". In: 19.6 (2020), pp. 1255–1279.

[7] Quan-Hoang Vo et al. "Multi-channel LSTM-CNN model for Vietnamese sentiment analysis". In: *2017 9th international conference on knowledge and systems engineering (KSE)*. IEEE. 2017, pp. 24–29.

[8] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[9] Mike Schuster and Kuldip K Paliwal. “Bidirectional recurrent neural networks”. In: *IEEE transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[10] Dat Quoc Nguyen and Anh Tuan Nguyen. “PhoBERT: Pre-trained language models for Vietnamese”. In: *Findings of EMNLP* (2020).

[11] Liu Zhuang et al. “A robustly optimized BERT pre-training approach with post-training”. In: *Proceedings of the 20th chinese national conference on computational linguistics*. 2021, pp. 1218–1227.

[12] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[13] Dat Quoc Nguyen and Anh Tuan Nguyen. “PhoBERT: Pre-trained language models for Vietnamese”. In: *arXiv preprint arXiv:2003.00744* (2020).

[14] Cu Vinh Loc et al. “A Text Classification for Vietnamese Feedback via PhoBERT-Based Deep Learning”. In: *Proceedings of Seventh International Congress on Information and Communication Technology: ICICT 2022, London, Volume 3*. Springer. 2022, pp. 259–272.

[15] Vu Cong Duy Hoang et al. “A comparative study on vietnamese text classification methods”. In: *2007 IEEE international conference on research, innovation and vision for the future*. IEEE. 2007, pp. 267–273.

[16] Son T Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. “Empirical study of text augmentation on social media text in vietnamese”. In: *arXiv preprint arXiv:2009.12319* (2020).

[17] Huu Nguyen Phat and Nguyen Thi Minh Anh. “Vietnamese text classification algorithm using long short term memory and Word2Vec”. In: 19.6 (2020), pp. 1255–1279.

[18] To Nguyen Phuoc Vinh and Ha Hoang Kha. “Vietnamese news articles classification using neural networks”. In: *Journal of Advances in Information Technology (JAIT)* (2021).

[19] Toan Pham Van and Ta Minh Thanh. “Vietnamese news classification based on BoW with keywords extraction and neural network”. In: *2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES)*. IEEE. 2017, pp. 43–48.

[20] Guojie Yang et al. “Interoperability and data storage in internet of multimedia things: investigating current trends, research challenges and future directions”. In: *IEEE Access* 8 (2020), pp. 124382–124401.