



INFO-H519 - NLP with Deep Learning - SP24

Fine Tuning BERT on Sentimental Analysis

BERT | LLM

Professor Ming Jiang

Group1

Hasaranga Jayathilake

Hari Shivani Gudi

Hymavathi Gummudala



LUDDY

SCHOOL OF INFORMATICS,
COMPUTING, AND ENGINEERING
INDIANAPOLIS

Lineup

- Introduction
 - Prior Related Work (Literature Review)
 - Data
 - Approach
 - Experiments
 - Results
 - Analysis & Conclusion



SECTION 1

Introduction

Background

- **Feedback** and **reviews** are important for the organizations to be relevant on the current market trends and achieve sustainable competitive advantage (Devlin, Chang, Lee, & Toutanova, 2018).
- **Sentiment analysis** using NLP is crucial for **leveraging feedback data effectively** and conduct market research in present of Big Data market.
- Predominately, Sentiment Analysis involves **classifying text** based on the sentiment or opinion expressed (**positive**, **negative**, or **neutral**).
- Developing sentiment analysis models for **multiple languages** allows us to better understand and engage with a **global audience**.



Languages Focused

Introduction

Vietnamese is spoken as first language by **approximately 86 million people** and features a complex phonetic and grammatical structure.

- There is less availability of pre-trained models and linguistic resources compared to English.
 - **There is an emerging interest in NLP research and application for Vietnamese.** (Bello, Ng, & Leung, 2023)

English is spoken as first language by around **1.45 billion people** and is the dominant language in trade, research and technology.

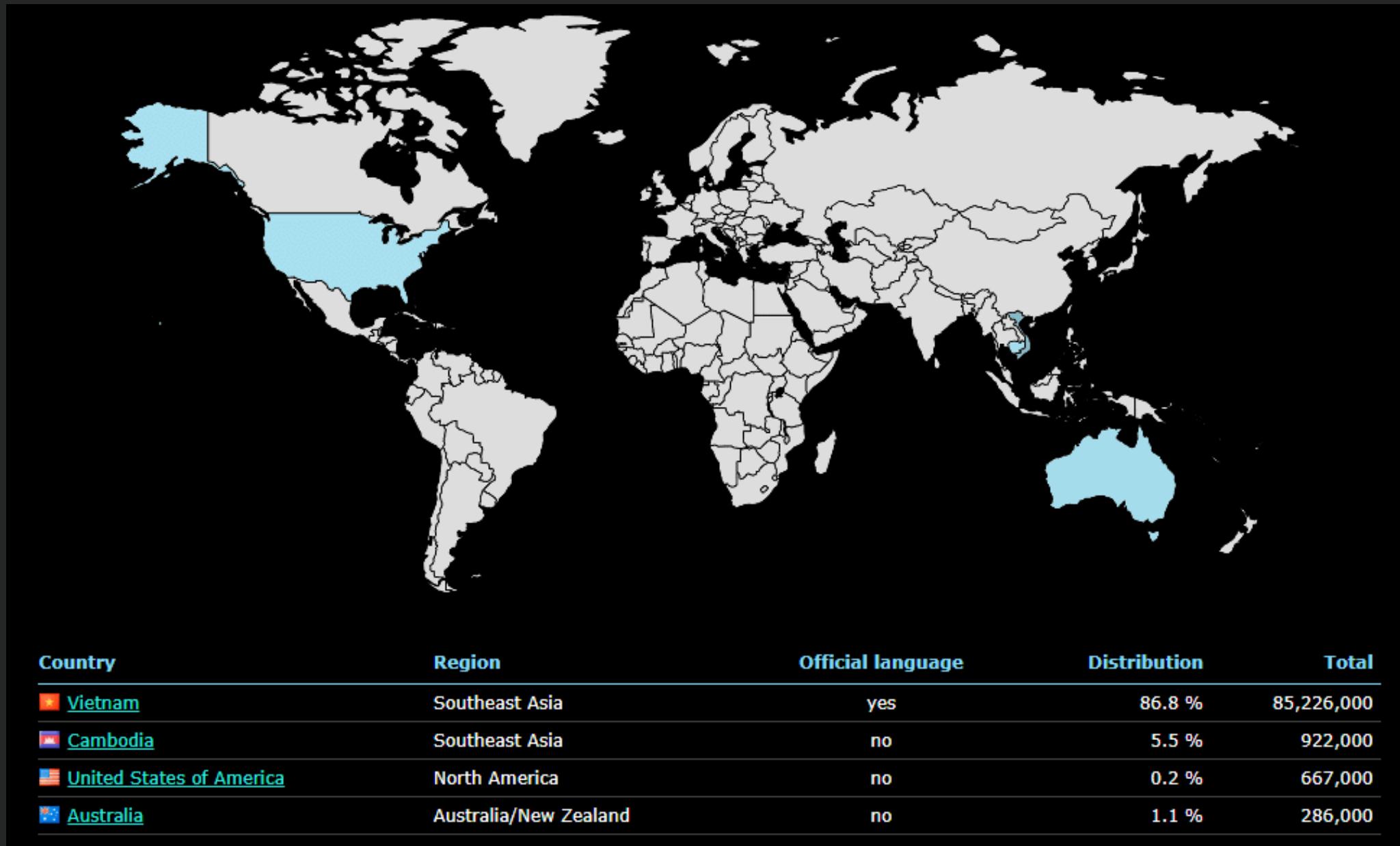
- Extensive pre-trained models and resources are available.
 - Well-established best practices and techniques for sentiment analysis.



Most Spoken Languages on Earth, 2023

Top 40 most spoken language in the world, based on the highest number of speakers

1	English	1.456 B	21	Vietnamese	86 M
2	Mandarin Chinese	1.138 B	22	Wu Chinese	83 M
3	Hindi	610 M	23	Tagalog	83 M
4	Spanish	559 M	24	Korean	82 M
5	French	310 M	25	Iranian Persian	79 M
6	Standard Arabic	274 M	26	Hausa	79 M
7	Bengali	273 M	27	Swahili	72 M
8	Portuguese	264 M	28	Javanese	68 M
9	Russian	255 M	29	Italian	68 M
10	Urdu	232 M	30	Western Punjabi	67 M
11	Indonesian	199 M	31	Gujarati	62 M
12	Standard German	133 M	32	Thai	61 M
13	Japanese	123 M	33	Kannada	59 M
14	Nigerian Pidgin	121 M	34	Amharic	58 M
15	Egyptian Arabic	102 M	35	Bhojpuri	52 M
16	Marathi	99 M	36	Eastern Punjabi	52 M
17	Telugu	96 M	37	Min Nan Chinese	50 M
18	Turkish	90 M	38	Jin Chinese	48 M
19	Tamil	87 M	39	Levantine Arabic	48 M
20	Yue Chinese	87 M	40	Yoruba	46 M



Objectives

Research Objectives

- Evaluate Sentiment detection accuracy on the current available models on Vietnamese Language

Model Name	Developer	GitHub Link
BERT-Base, Multilingual Cased	Google (Devlin et al. In 2018)	Link
Fine-Tune BERT-Base, Multilingual Cased	Nguyen, et al., 2020	Link
PhoBERT Model	Nguyen, et al., 2020	Link

- Fine-Tune by combining current models with other model architectures.

Model Name	New Mode Integrations
BERT-Base, Multilingual Cased	+ LSTM, +FastText, + Glove
Fine-Tune BERT-Base, Multilingual Cased	+ LSTM, +FastText, + Glove
PhoBERT Model	+ LSTM, +FastText, + Glove

- Evaluate and Compare the model performances on English language dataset

Dataset Name	Link
Vietnamese Hotel Reviews (NTC-SV)	Link
English IMDB Movie Reviews	Link

Evaluation Parameters
Accuracy (%)
Precision (%)
Recall (%)
F1 (%)

- Provide insights into the challenges and opportunities of conducting sentiment analysis across multiple languages.

- Evaluate Sentiment detection accuracy on the current available models on Vietnamese Language

Model Name	Developer	GitHub Link
BERT-Base, Multilingual Cased	Google (Devlin et al. in 2018)	Link
Fine-Tune BERT-Base, Multilingual Cased	Nguyen, et al., 2020	Link
PhoBERT Model	Nguyen, et al., 2020	Link

- Fine-Tune by combining current models with other model architectures.

Model Name	New Mode Integrations
BERT-Base, Multilingual Cased	+ LSTM, +FastText, + Glove
Fine-Tune BERT-Base, Multilingual Cased	+ LSTM, +FastText, + Glove
PhoBERT Model	+ LSTM, +FastText, + Glove

- Evaluate and Compare the model performances on English language dataset

Dataset Name	Link
Vietnamese Hotel Reviews (NTC-SV)	Link
English IMDB Movie Reviews	Link

Evaluation Parameters
Accuracy (%)
Precision (%)
Recall (%)
F1 (%)

- Provide insights into the challenges and opportunities of conducting sentiment analysis across multiple languages.

Section 2

Prior Related Work

Literature Review

Literature Review

Focusing on sentiment analysis and the fine-tuning of BERT models

Development of BERT and Its Impact

- Jacob et al. (2018) BERT revolutionized NLP by understanding context in text efficiently.
- BERT's introduction changed how models process language, improving understanding across various NLP applications (Dang et al., 2022).

Research Paper Name and Reference	Methods Used
Multi-task solution for aspect category sentiment analysis on Vietnamese datasets (Dang et al., 2022).	Transfer learning on BERT model

Multilingual and Cross-Lingual Challenges

- BERT adaptations handle multiple languages, aiding sentiment analysis where resources for languages like Vietnamese lag (Lê et al., 2020).
- Multilingual BERT overcomes resource gaps in languages, enhancing sentiment analysis across diverse linguistic landscapes (Lê et al., 2020).

Research Paper Name and Reference	Methods Used
On Vietnamese sentiment analysis: A transfer learning method. (Lê et al., 2020).	Multilingual pre-trained language BERT model

Integration of BERT with Other Models

- Integrating BERT with LSTM improves sentiment analysis by enhancing text classification capabilities significantly (Luc, et al, 2021).
- Studies show BERT, combined with deep learning models like CNNs, elevates performance in text analysis. (Luc, et al, 2021).

Research Paper Name and Reference	Methods Used
From aspect-based sentiment analysis to social listening system for business intelligence. (Luc, et al, 2021).	Ensemble deep learning architecture CNN and Bi-LSTM models

Effectiveness of Fine-Tuning on Domain-Specific Datasets

- Fine-tuning BERT on domain-specific datasets significantly boosts performance by aligning model focus with contextual nuances (Thin, et al., 2020).
- Howard et al. demonstrated that domain-specific fine-tuning of BERT enhances model accuracy and relevance (Thin, et al., 2020).

Research Paper Name and Reference	Methods Used
Two new large corpora for Vietnamese aspect-based sentiment analysis at sentence level. (Thin, et al., 2020).	Fine-tuning the BERT-based models.



LUDDY

School of Informatics, Computing, and Engineering Indianapolis

Development of BERT and Its Impact

- Jacob et al. (2018) BERT revolutionized NLP by understanding context in text efficiently.
- BERT's introduction changed how models process language, improving understanding across various NLP applications (Dang et al., 2022).

Research Paper Name and Reference

Multi-task solution for aspect category sentiment analysis on Vietnamese datasets (Dang et al., 2022).

Methods Used

Transfer learning on BERT model

Multilingual and Cross-Lingual Challenges

- BERT adaptations handle multiple languages, aiding sentiment analysis where resources for languages like Vietnamese lag (Lê et al., 2020).
- Multilingual BERT overcomes resource gaps in languages, enhancing sentiment analysis across diverse linguistic landscapes (Lê et al., 2020).

Research Paper Name and Reference

On Vietnamese sentiment analysis: A transfer learning method. (Lê et al., 2020).

Methods Used

Multilingual pre-trained language BERT model



LUDDY

School of Informatics, Computing, and Engineering Indianapolis

Integration of BERT with Other Models

- Integrating BERT with LSTM improves sentiment analysis by enhancing text classification capabilities significantly (Luc, et al, 2021).
- Studies show BERT, combined with deep learning models like CNNs, elevates performance in text analysis. (Luc, et al, 2021).

Research Paper Name and Reference	Methods Used
From aspect-based sentiment analysis to social listening system for business intelligence. (Luc, et al, 2021).	Ensemble deep learning architecture CNN and Bi-LSTM models

Effectiveness of Fine-Tuning on Domain-Specific Datasets

1. Fine-tuning BERT on domain-specific datasets significantly boosts performance by aligning model focus with contextual nuances (Thin, et al., 2020).
2. Howard et al. demonstrated that domain-specific fine-tuning of BERT enhances model accuracy and relevance (Thin, et al., 2020).

Research Paper Name and Reference

Two new large corpora for Vietnamese aspect-based sentiment analysis at sentence level. (Thin, et al., 2020).

Methods Used

Fine-tuning the BERT-based models.

Limitations and Gaps in Existing Research

Lack of comparison across models specialized for Vietnamese language (e.g., PhoBERT)

- Limited direct comparisons of Vietnamese-specific models (e.g., PhoBERT) hinder understanding of their effectiveness.
- Few studies analyze how different models handle Vietnamese language nuances.

Limited evaluation of model effectiveness for Vietnamese language analysis

- Lack of broad assessment across a range of metrics for Vietnamese datasets.
- Insufficient analysis of model performance in different domains.

(Van Thinh, Hao, & Nguyen, 2020)

Lack comparison with other languages like English to measure generalizability

(Luong Luc Phan et al., 2021)

- Limited comparisons with other languages (e.g., English) restrict the measurement of model generalizability.
- Lack of cross-language insights hampers understanding of model adaptability and robustness.



LUDDY

School of Informatics, Computing, and Engineering Indianapolis

(Luong Luc Phan et al., 2021)

Lack of comparison across models specialized for Vietnamese language (e.g., PhoBERT)

- Limited direct comparisons of Vietnamese-specific models (e.g. PhoBERT) hinder understanding of their effectiveness.
- Few studies analyze how different models handle Vietnamese language nuances.

Limited evaluation of model effectiveness for Vietnamese language analysis

- Lack of broad assessment across a range of metrics for Vietnamese datasets.
- Insufficient analysis of model performance in different domains.



LUDDY

School of Informatics, Computing, and Engineering Indianapolis

(Luong Luc Phan et al., 2021)

Lack comparison with other languages like English to measure generalizability

- Limited comparisons with other languages (e.g., English) restrict the measurement of model generalizability.
- Lack of cross-language insights hampers understanding of model adaptability and robustness.



Project Focus: (Novelty of the Project)

- Research explores **Vietnamese models** for **sentiment analysis on English texts**.
- Aims to assess **model effectiveness in machine translation and paraphrasing**.
- **There is possibility to Vietnamese models may outperform multilingual models in language tasks (Hypothesis of Project)**
- Study investigates if these advantages improve English machine translation.

Section 3

Data

Datasets

IMDB Dataset

Dataset Analogy	
Movie Review descriptions	
50000 Rows of Text Data	
Sentiment defined using 2 Labels	
Label: 0 - Negative	
Label: 1 - Positive	

		text	label
0	Forget what I said about Emeril. Rachael Ray i...	0	
1	Former private eye-turned-security guard ditch...	0	
2	Mann photographs the Alberta Rocky Mountains i...	0	
3	Simply put: the movie is boring. Cliché upon c...	0	
4	Now being a fan of sci fi, the trailer for thi...	1	
...	
49995	The "documentary", and we use that term loosel...	0	
49996	This outlandish Troma movie is actually a very...	1	
49997	I found the film Don't Look In The Basement to...	1	
49998	I have read the novel Reaper of Ben Mezrich a ...	0	
49999	Went to see this finnish film and I've got to ...	1	
[50000 rows x 2 columns]			

Vietnamese Dataset

Dataset Analogy	
Pre-existing split dataset with train	
and test sub-datasets	
Hotel Food Review Descriptions	
Train Dataset	Test Dataset
40761 Rows Data	10000 Rows Data
Sentiment defined using 2 Labels	
Label: 0 - Negative	
Label: 1 - Positive	

		text	label
0	dồ_ăn ngon positive hợp_khẩu vị nhiều món nhâ... n	0	
1	chè bd thdm positive có vị ngọt ngọt nhưng lại... n	0	
2	chiều hôm nay mới đi ăn về nghe thiên_hạ đồn q... n	0	
3	mình đặt_hàng qua tin nhắn với cửa_hàng hứa sá... n	0	
4	ghé mấy lần rồi mà không review đi đâu cũng ch... n	1	

IMDB Dataset

Dataset Analogy

Movie Review descriptions

50000 Rows of Text Data

Sentiment defined using 2 Labels

Label: 0 - Negative

Label: 1 - Positive

	text	label
0	Forget what I said about Emeril. Rachael Ray i...	0
1	Former private eye-turned-security guard ditch...	0
2	Mann photographs the Alberta Rocky Mountains i...	0
3	Simply put: the movie is boring. Cliché upon c...	0
4	Now being a fan of sci fi, the trailer for thi...	1
...
49995	The "documentary", and we use that term loosel...	0
49996	This outlandish Troma movie is actually a very...	1
49997	I found the film Don't Look In The Basement to...	1
49998	I have read the novel Reaper of Ben Mezrich a ...	0
49999	Went to see this finnish film and I've got to ...	1

[50000 rows x 2 columns]



LUDDY

School of Informatics, Computing, and Engineering Indianapolis

Vietnamese Dataset

Dataset Analogy

Pre-existing split dataset with train and test sub-datasets

Hotel Food Review Descriptions

Train Dataset

40761 Rows Data

Sentiment defined using 2 Labels

Label: 0 - Negative

Label: 1 - Positive

	text	label
0	đồ_ăn ngon positive hợp_khẩu vị nhiều món nhân...	0
1	chè bơ thơm positive có vị ngọt ngọt nhưng lại...	0
2	chiều hôm nay mới đi ăn về nghe thiên_hạ đồn q...	0
3	mình đặt_hàng qua tin nhắn với cửa_hàng hứa sá...	0
4	ghé mấy lần rồi mà không review đi đâu cũng ch...	1

Data Distribution across Different Labels

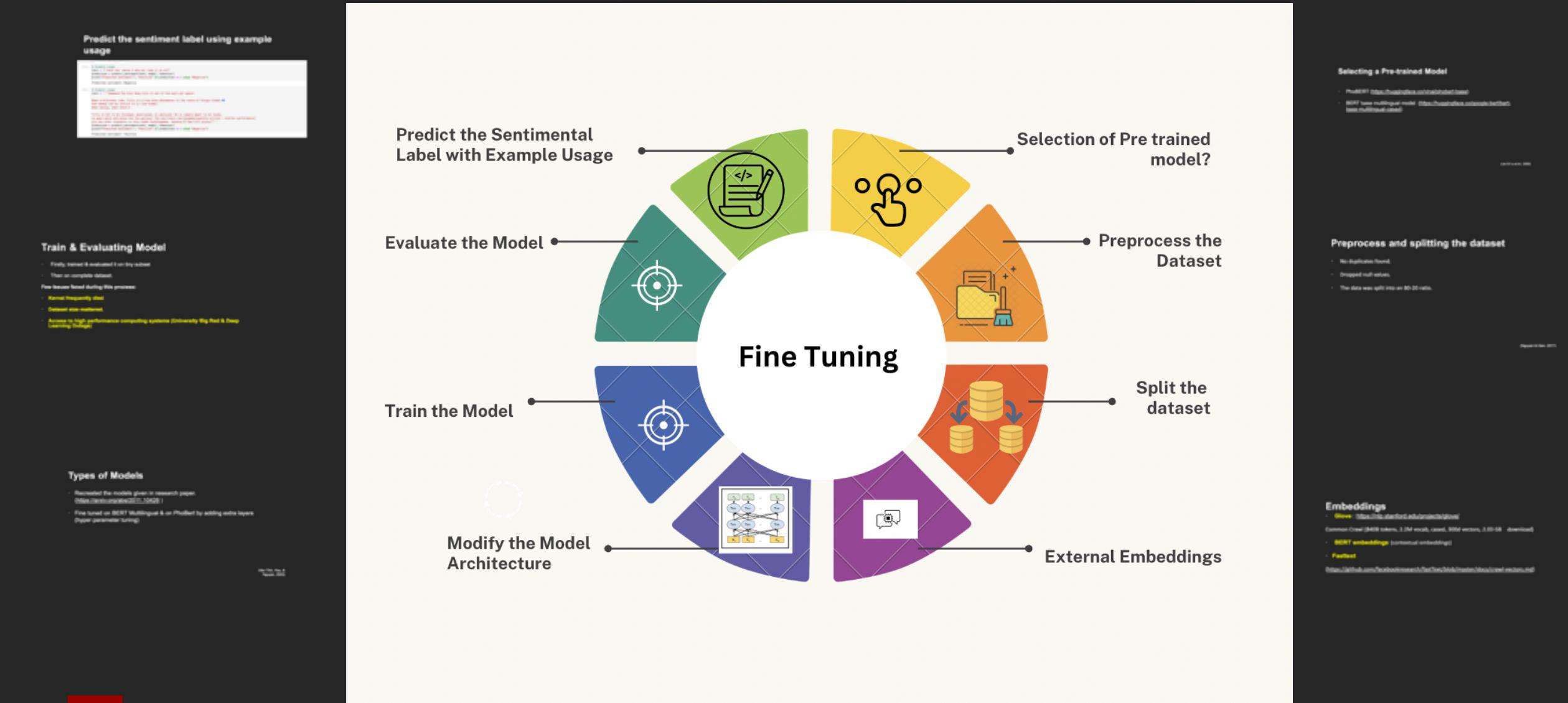
Dataset	Train		Test	
	Positive	Negative	Positive	Negative
IMDB	19961	20039	5039	4961
Vietnamese	20493	20268	5000	5000

Section 4

Approach

Fine Tuning

Approach



Selecting a Pre-trained Model

- PhoBERT (<https://huggingface.co/vinai/phobert-base>)
- BERT base multilingual model (<https://huggingface.co/google-bert/bert-base-multilingual-cased>)

Preprocess and splitting the dataset

- No duplicates found.
- Dropped null values.
- The data was split into an 80-20 ratio.

Embeddings

- **Glove** : <https://nlp.stanford.edu/projects/glove/>

Common Crawl (840B tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB download)

- **BERT embeddings** (contextual embeddings)
- **Fasttext**

(<https://github.com/facebookresearch/fastText/blob/master/docs/crawl-vectors.md>)

Types of Models

- Recreated the models given in research paper.
(<https://arxiv.org/abs/2011.10426>)
- Fine tuned on BERT Multilingual & on PhoBert by adding extra layers
(hyper parameter tuning)

Train & Evaluating Model

- Firstly, trained & evaluated it on tiny subset
- Then on complete dataset.

Few Issues faced during this process:

- **Kernel frequently died**
- **Dataset size mattered.**
- **Access to high performance computing systems (University Big Red & Deep Learning Outage)**

Predict the sentiment label using example usage

```
[21]: # Example usage
text = "I hate the movie I did not like it at all"
prediction = predict_sentiment(text, model, tokenizer)
print("Predicted sentiment:", "Positive" if prediction == 1 else "Negative")
```

Predicted sentiment: Negative

```
[22]: # Example usage
text = """Aaaaand The Star Buoy hits it out of the park yet again!
```

What a hilarious ride. Tillu is a true blue phenomenon in the realm of Telugu Cinema ❤
And nobody can do justice to it like Siddu!
What energy, what charm ❤

Tillu is not to be reviewed, questioned, or analyzed. He is simply meant to be loved,
so gooo watch and enjoy the fun partyyy! The one-liners and Anupama(superbly written - stellar performance)
are the other standouts in this Siddu Jonnalagadda bonanza 🎉 Don'ttt missss!""""

```
prediction = predict_sentiment(text, model, tokenizer)
print("Predicted sentiment:", "Positive" if prediction == 1 else "Negative")
```

Predicted sentiment: Positive



LUDDY

School of Informatics, Computing, and Engineering Indianapolis

Section 5

Experiments

Experiments

Experiments

Experimental Design and Setup

- Preprocessing Included Removal:** All models used data cleaning, such as removing duplicates.
- Model Selection Criteria:** Models were chosen based on their language compatibility and performance benchmarks.
- Training Protocols Established:** Training involved multiple epochs, utilizing cross-validation to ensure robustness.

(Bello, Ng, & Leung, 2020)

Datasets Used

- ❖ **Diverse Dataset Utilization:** Models were trained on datasets like IMDB reviews and Vietnamese hotel reviews.
- ❖ **Validation Methods:** Employed cross-validation techniques to assess model generalization across different datasets.

Dataset	Train		Test	
	Positive	Negative	Positive	Negative
IMDB	19981	20039	5039	4961
Vietnamese	20493	20268	5000	5000

(Bello, Ng, & Leung, 2020)

Training Process

- Use of High-Performance Computing:** Models required substantial computational resources for training.
- Challenges in Training:** Encountered issues like kernel crashes due to intensive computation demands.
- Training Duration:** Training times varied, with adjustments made based on the preliminary results and computational limits.

(Luu & Le et al., 2020)

Evaluation Metrics

- Accuracy and Precision:** Metrics focused on how precisely models predicted sentiment categories.
- Recall and F1 Score:** These metrics helped evaluate the models' ability to identify all relevant instances.
- Consistent Metric Application:** All models were evaluated using the same criteria for fair comparison.

(Luu & Le et al., 2020)

Baseline Comparisons

- Comparison Against Previous Models:** Baseline models included earlier versions of BERT and other NLP frameworks.
- Benchmarking Against Industry Standards:** Compared results with published benchmarks to validate improvements.
- Performance Enhancements Noted:** Detailed analysis of how fine-tuning and model adjustments outperformed baselines.

(Van Thien, Hieu, & Nguyen, 2020)

Experimental Design and Setup

1. **Preprocessing Included Removal:** All models used data cleaning, such as removing duplicates.
2. **Model Selection Criteria:** Models were chosen based on their language compatibility and performance benchmarks.
3. **Training Protocols Established:** Training involved multiple epochs, utilizing cross-validation to ensure robustness.

Datasets Used

- ❖ **Diverse Dataset Utilization:** Models were trained on datasets like IMDB reviews and Vietnamese hotel reviews.
- ❖ **Validation Methods:** Employed cross-validation techniques to assess model generalization across different datasets.

Dataset	Train		Test	
	Positive	Negative	Positive	Negative
IMDB	19961	20039	5039	4961
Vietnamese	20493	20268	5000	5000

Model Configurations

1. **Base and Fine-Tuned Models:** Both standard and fine-tuned versions of BERT were used to compare performance.
2. **Integration with LSTM:** Models incorporated LSTM layers to enhance learning sequential patterns in text.
3. **Architecture Specifics Detailed:** Configurations included adjustments in layers and parameters for optimal training.

Training Process

1. **Use of High-Performance Computing:** Models required substantial computational resources for training.
2. **Challenges in Training:** Encountered issues like kernel crashes due to intensive computation demands.
3. **Training Duration:** Training times varied, with adjustments made based on the preliminary results and computational limits.

Evaluation Metrics

1. **Accuracy and Precision:** Metrics focused on how precisely models predicted sentiment categories.
2. **Recall and F1 Score:** These metrics helped evaluate the models' ability to identify all relevant instances.
3. **Consistent Metric Application:** All models were evaluated using the same criteria for fair comparison.

Baseline Comparisons

1. **Comparison Against Previous Models:** Baseline models included earlier versions of BERT and other NLP frameworks.
2. **Benchmarking Against Industry Standards:** Compared results with published benchmarks to validate improvements.
3. **Performance Enhancements Noted:** Detailed analysis of how fine-tuning and model adjustments outperformed baselines.



Section 6

Results

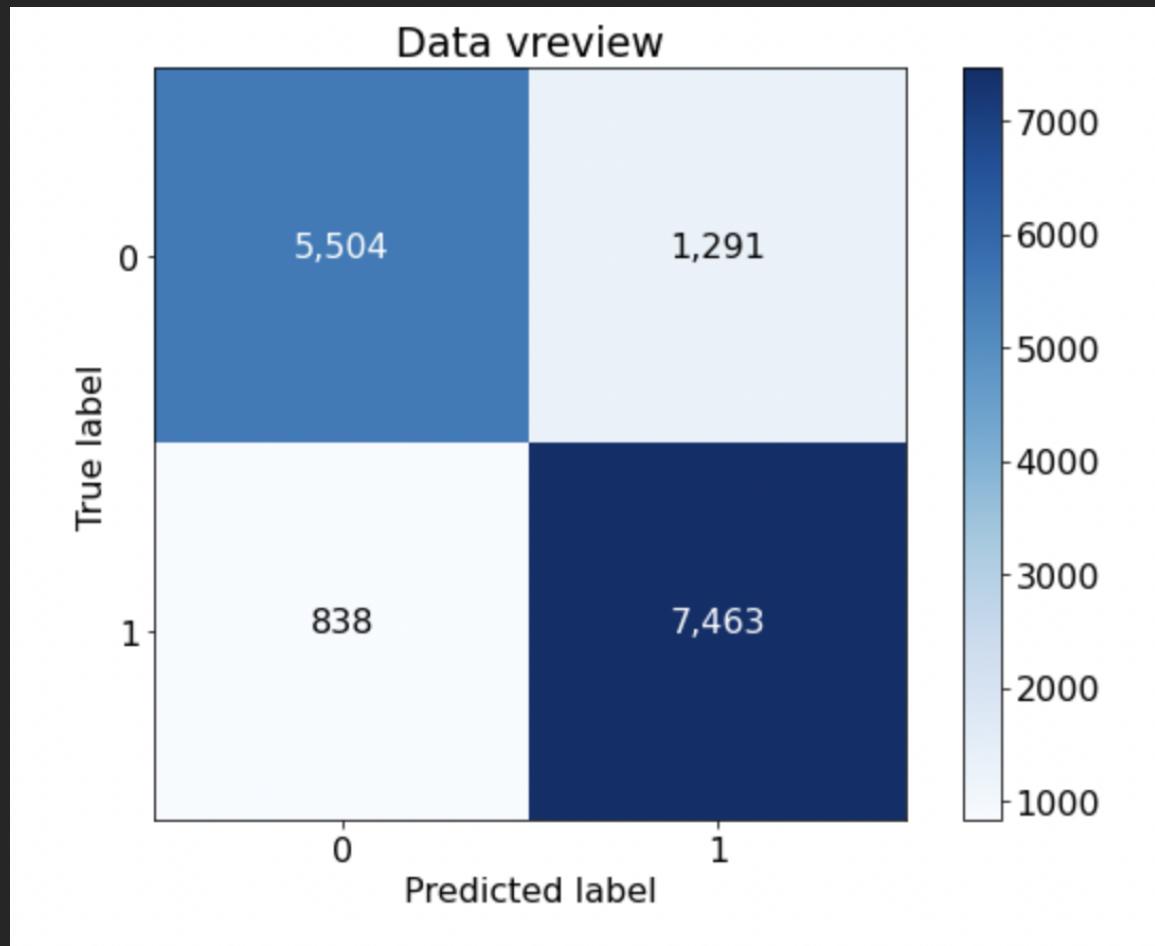
Fine-Tune BERT-Base, Multilingual Cased

TABLE 5. RESULT OF OUR MODEL ON NTC-SV DATASET COMPARED TO OTHER MODELS

Model	Precision(%)	Recall(%)	F1(%)
SVM	89.23	92.52	90.84
XGBoost	88.76	90.58	89.63
FastText + TextCNN	67.9	89.1	77.1
FastText + LSTM	88.5	89.7	89.1
FastText + RCNN	89.2	91.7	90.4
Glove + TextCNN	69.7	87.7	77.7
Glove + LSTM	88.7	91.8	89.8
Glove + RCNN	85.8	85.8	90.7
BERT-base	88.13	94.02	90.9
BERT-LSTM	89.78	92.08	90.91
BERT-TextCNN	88.85	93.14	90.94
BERT-RCNN	88.76	93.68	91.15

Models (epoch = 5)	Datasets			
	Vietnamese			
	Accuracy	Precision	Recall	F1
Fine-Tune BERT- Base, Multilingua l Cased	0.8296	0.8097	0.8618	0.8349
Fine-Tune BERT- Base, Multilingua l Cased + LSTM	0.8515	0.7737	0.8547	0.8122

BERT-Base, Multilingual Cased



Models (epoch = 5)	Datasets			
	Vietnamese			
	Accuracy	Precision	Recall	F1
BERT- Base, Multilingua l Cased	0.8932	0.8795	0.9112	0.8951
BERT- Base, Multilingua l Cased + LSTM	0.7564	0.7459	0.7806	0.7629

Fine-Tune BERT-Base, Multilingual Cased

Original PhoBERT-base Results:

Accuracy: 96.7%

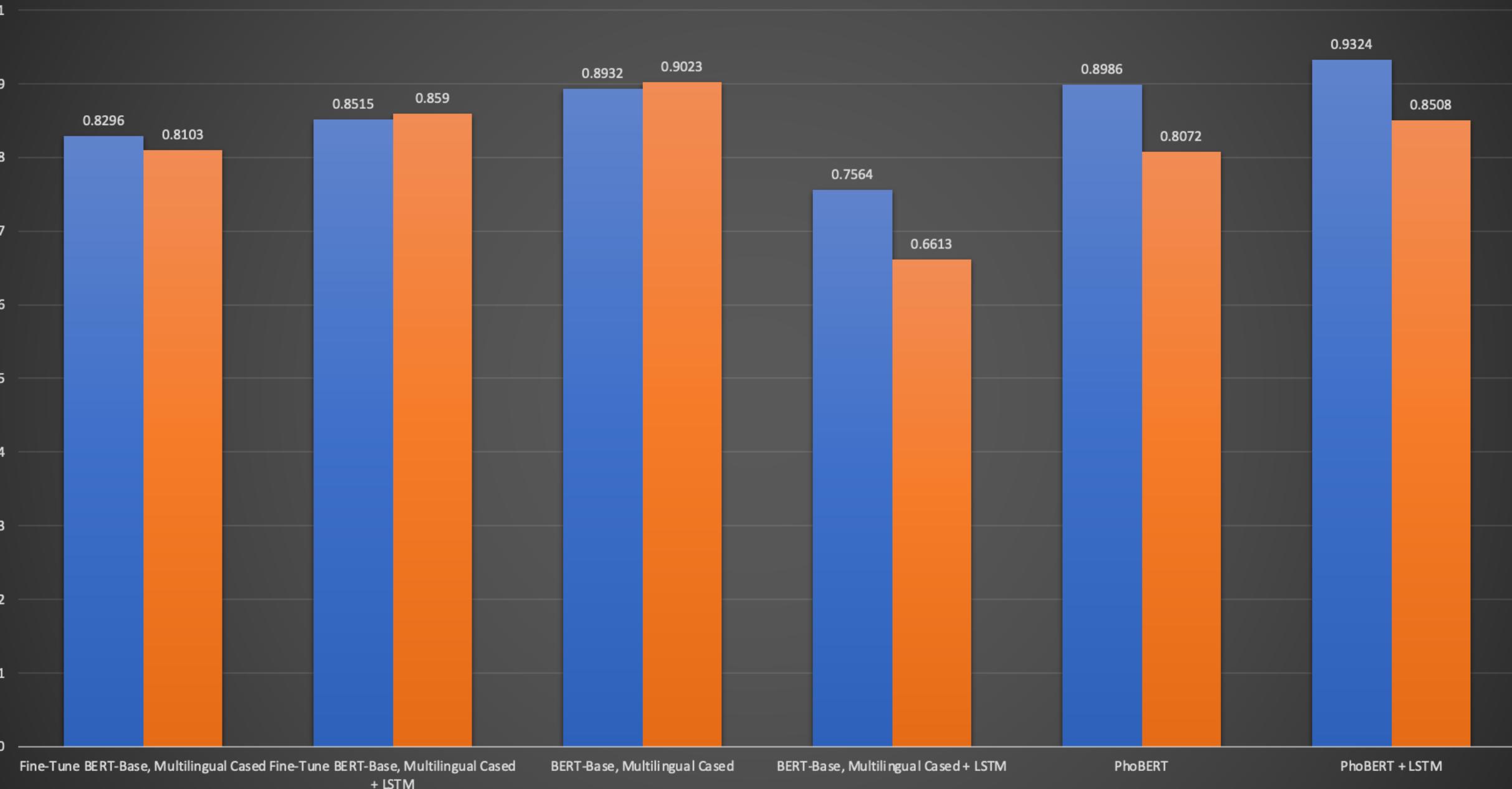
F1 Score: 93.6%

Models (epoch = 5)	Datasets			
	Vietnamese			
	Accuracy	Precision	Recall	F1
PhoBERT	0.8986	0.9008	0.9394	0.9197
PhoBERT + LSTM	0.9324	0.8959	0.9532	0.9236

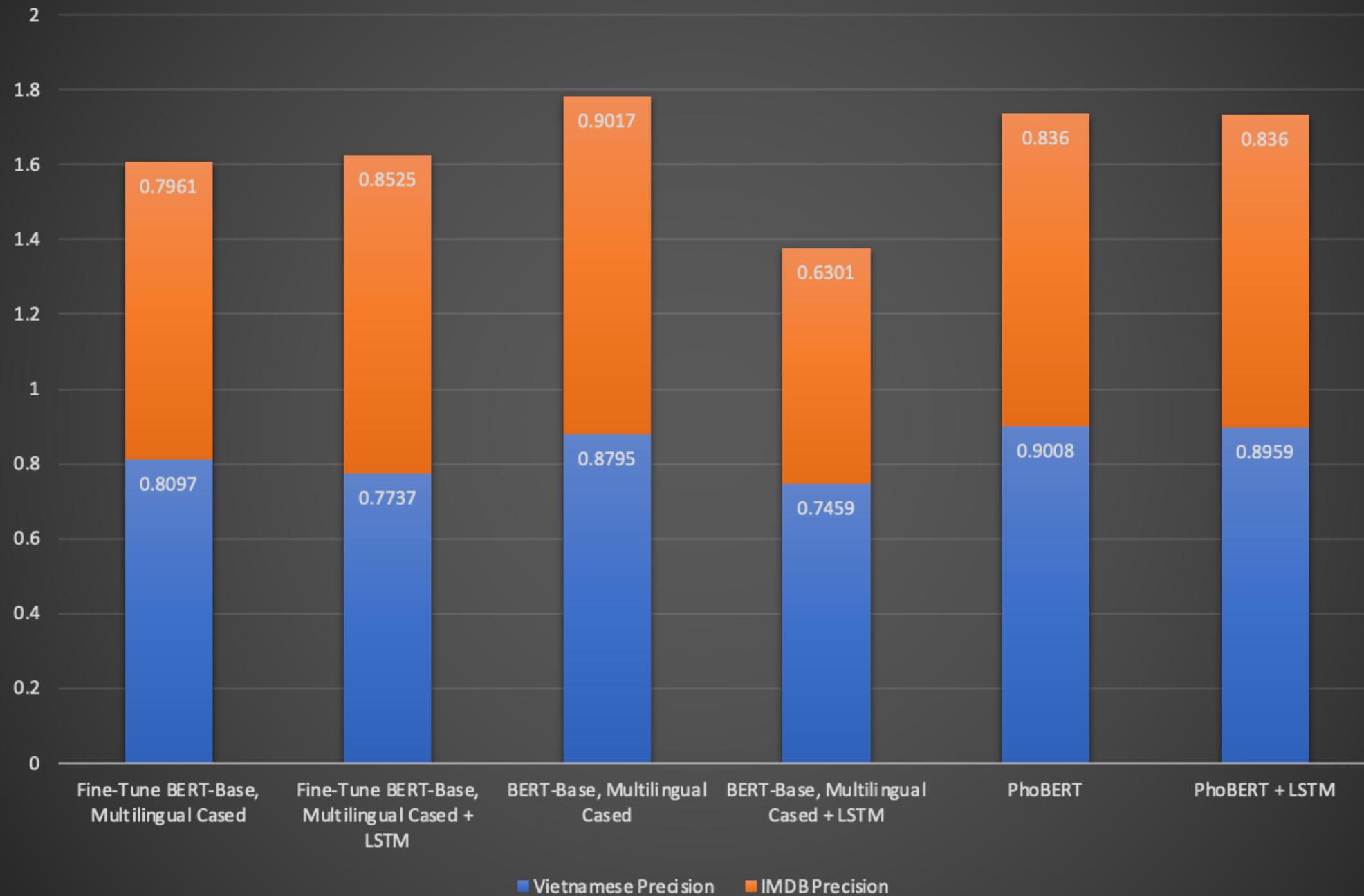
Models (epoch = 5)	Datasets							
	Vietnamese				English			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Fine-Tune BERT-Base, Multilingual Cased	0.8296	0.8097	0.8618	0.8349	0.8103	0.7961	0.8383	0.8166
Fine-Tune BERT-Base, Multilingual Cased + LSTM	0.8515	0.7737	0.8547	0.8122	0.859	0.8525	0.8990	0.8757
BERT-Base, Multilingual Cased	0.8932	0.8795	0.9112	0.8951	0.9023	0.9017	0.9047	0.9032
BERT-Base, Multilingual Cased + LSTM	0.7564	0.7459	0.7806	0.7629	0.6613	0.6301	0.7735	0.6945
PhoBERT	0.8986	0.9008	0.9394	0.9197	0.8072	0.836	0.8091	0.8223
PhoBERT + LSTM	0.9324	0.8959	0.9532	0.9236	0.8508	0.836	0.8277	0.8319

Phobert Vs Vietnamese Accuracy

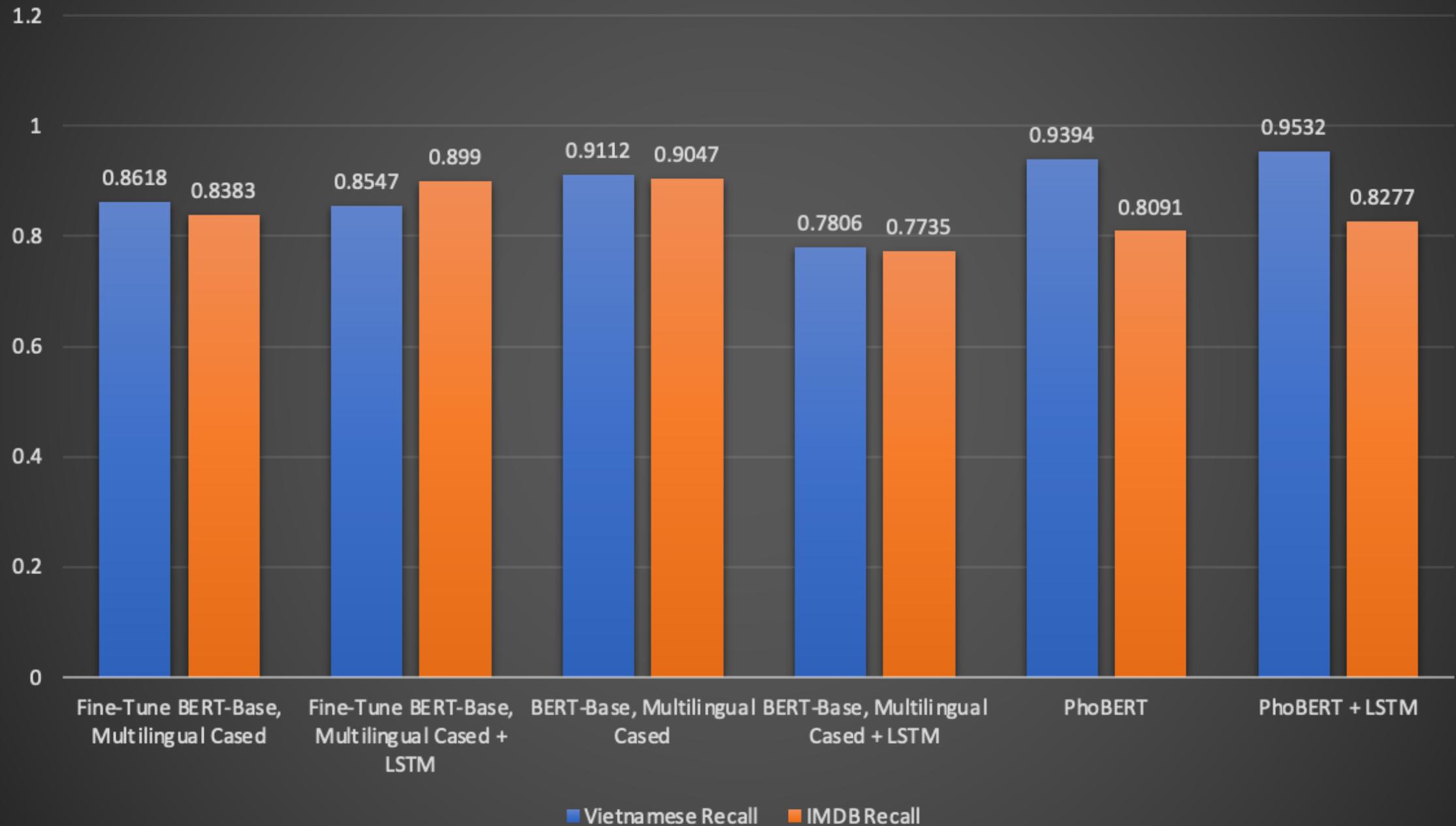
Vietnamese Accuracy IMDB Accuracy



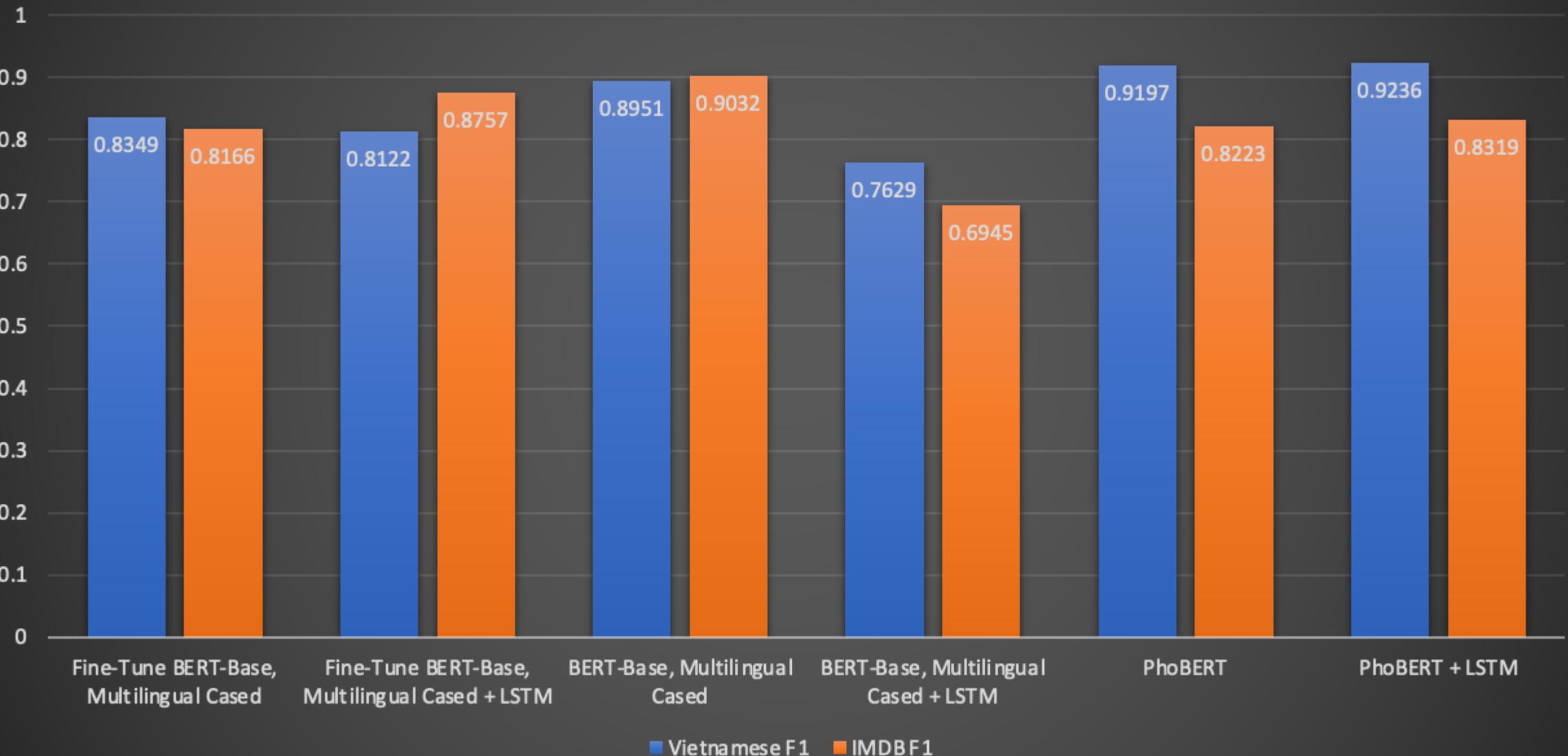
Precision



Recall



F1 Score

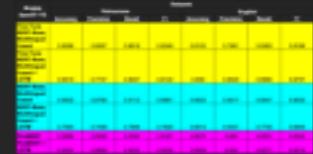


Section 7

Analysis & Conclusion

Analysis & Conclusion

Analysis & Conclusion



Performance Analysis

- PhoBERT models outperform Google BERT on both English and Vietnamese datasets.
- LSTM integration improves recall but may slightly decrease precision in some models.
- PhoBERT with LSTM shows the best overall performance, particularly in F1 score.



Model Effectiveness

- PhoBERT is more effective across languages, showcasing strong cross-lingual capabilities.
- BERT models demonstrate high recall, indicating better capture of relevant sentiment.
- The addition of LSTM consistently enhances sentiment detection in diverse datasets.



Theoretical Implications

- Identified that Vietnamese models have better efficacy compared to the BERT multi-model in sentiment analysis, using only 5 epochs.
- The results affirm the importance of language-specific models for sentiment analysis tasks.
- The effectiveness of LSTM implies sequential data processing is critical for sentiment analysis.
- The strong performance of PhoBERT suggests that contextualized embeddings play a key role in NLP.



Practical Applications

- PhoBERT could be used in customer service bots for sentiment recognition across languages.
- Models with LSTM can improve sentiment-based recommendation systems in e-commerce.
- Fine-tuned BERT models are applicable for social media monitoring in multiple languages.



Challenges and Limitations

Model Name	Original Optimization Parameters			In this Project Optimization Activities		
	Epochs	Max Length	Params	Epochs	Max Length	Params
BERT-Base, Multilingual Cased	40	512	179M	5	512	179M
Fine-Tune BERT-Base, Multilingual Cased	10	256	179M	5	256	179M
PhoBERT	40	256	135M	5	256	135M

- Multilingual BERT underperforms compared to monolingual PhoBERT in language-specific tasks.
- Integrating LSTM shows varied results, suggesting potential overfitting or bias in certain contexts.
- Computational constraints may limit the extent of fine-tuning and model complexity.

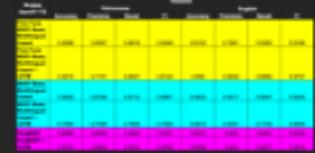
(Nguyen et al., 2020)

Future Work and Improvements

- Further exploration of hyperparameter tuning could enhance model accuracies.
- Investigate models' performance on a wider range of languages and domains.
- Develop lightweight models to address computational constraints and enhance accessibility.
- As the hypothesis examined from the project needs to be investigated with very high iterations to understand the performance between Vietnamese-trained models versus multi-model efficacy.

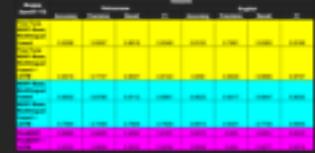
(Nguyen & Arora, 2021)

Performance Analysis



1. **PhoBERT models outperform** Google BERT on both English and Vietnamese datasets.
2. LSTM integration improves recall but may slightly decrease precision in some models.
3. PhoBERT with LSTM shows the best overall performance, particularly in F1 score.

Model Effectiveness



1. **PhoBERT is more effective across languages, showcasing strong cross-lingual capabilities.**
2. BERT models demonstrate high recall, indicating better capture of relevant sentiment.
3. The addition of LSTM consistently enhances sentiment detection in diverse datasets.



Challenges and Limitations

Model Name	Original Optimization Parameters			In this Project Optimization Activities		
	Epochs	Max Length	#params	Epochs	Max Length	#params
BERT-Base, Multilingual Cased	40	512	179M	5	512	179M
Fine-Tune BERT-Base, Multilingual Cased	10	256	179M	5	256	179M
PhoBERT	40	256	135M	5	256	135M

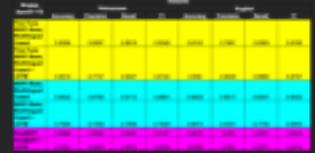
1. Multilingual BERT underperforms compared to monolingual PhoBERT in language-specific tasks.
2. Integrating LSTM shows varied results, suggesting potential overfitting or bias in certain contexts.
3. Computational constraints may limit the extent of fine-tuning and model complexity.

Theoretical Implications



1. Identified that **Vietnamese models have better efficacy compared to the BERT multi-model** in sentiment analysis, **using only 5 epochs**.
2. The results affirm the importance of language-specific models for sentiment analysis tasks.
3. The effectiveness of LSTM implies sequential data processing is critical for sentiment analysis.
4. **The strong performance of PhoBERT suggests that contextualized embeddings play a key role in NLP.**

Practical Applications



1. **PhoBERT** could be used in **customer service bots for sentiment recognition across languages**.
2. **Models with LSTM** can improve **sentiment-based recommendation systems in e-commerce**.
3. **Fine-tuned BERT models** are applicable for **social media monitoring in multiple languages**.

Future Work and Improvements

1. Further exploration of **hyperparameter tuning could enhance model accuracies.**
2. **Investigate models' performance on a wider range of languages and domains.**
3. **Develop lightweight models** to address computational constraints and enhance accessibility.
4. As the hypothesis examined from the project **needs to be investigated with very high iterations to understand the performance between Vietnamese-trained models versus multi-model efficacy.**



Reference

Bello, A., Ng, S. C., & Leung, M. F. (2023). A BERT framework to sentiment analysis of tweets. *Sensors*, 23(1), 506. <https://doi.org/10.3390/s23010506>

Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y. A., Gelbukh, A., & Zhou, Q. (2016). Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Cognitive Computation*, 8(4), 757–771. <https://doi.org/10.1007/s12559-016-9415-7>

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. <https://arxiv.org/abs/1810.04805>

Dang Van Thin, Duong Ngoc Hao, Vu Xuan Hoang, & Ngan Luu-Thuy Nguyen. (2022). Investigating monolingual and multilingual BERT models for Vietnamese aspect category detection. In Proceedings of the RIVF International Conference on Computing and Communication Technologies (RIVF'22) (pp. 130–135). IEEE. <https://doi.org/10.1109/RIVF55975.2022.10013792>



Reference

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv. <https://doi.org/10.48550/arXiv.1801.06146>

Lac Si Le, Dang Van Thin, Ngan Luu-Thuy Nguyen, & Son Quoc Trinh. (2020). A multi-filter BiLSTM-CNN architecture for Vietnamese sentiment analysis. In M. Hernes, K. Wojtkiewicz, & E. Szczerbicki (Eds.), Advances in Computational Collective Intelligence (pp. 752–763). Springer International Publishing.

Luong Luc Phan, Phuc Huynh Pham, Kim Thi-Thanh Nguyen, Sieu Khai Huynh, Tham Thi Nguyen, Luan Thanh Nguyen, Tin Van Huynh, & Kiet Van Nguyen. (2021). SA2SL: From aspect-based sentiment analysis to social listening system for business intelligence. In H. Qiu, C. Zhang, Z. Fei, M. Qiu, & S.-Y. Kung (Eds.), Knowledge Science, Engineering and Management (pp. 647–658). Springer International Publishing.

Manias, G., Mavrogiorgou, A., Kiourtis, A., Symvoulidis, C., & Kyriazis, D. (2023). Multilingual text categorization and sentiment analysis: A comparative analysis of the utilization of multilingual approaches for classifying twitter data. *Neural Computing and Applications*, 35(29), 21415–21431. <https://doi.org/10.1007/s00521-023-08629-3>



Reference

Nguyen, L. T., & Dien, D. (2017). English-Vietnamese cross-language paraphrase identification method. ACM International Conference Proceeding Series, 2017-December, 42–49.
<https://doi.org/10.1145/3155133.3155187>

Nguyen, Q. T., Nguyen, T. L., Luong, N. H., & Ngo, Q. H. (2020). Fine-tuning BERT for sentiment analysis of Vietnamese reviews. In Proceedings - 2020 7th NAFOSTED Conference on Information and Computer Science (NICS 2020) (pp. 302–307). IEEE. <https://doi.org/10.1109/NICS51282.2020.9335899>

Van Thin, D., Hao, D. N., & Nguyen, N. L. T. (2023). Vietnamese sentiment analysis: An overview and comparative study of fine-tuning pretrained language models. ACM Transactions on Asian and Low-Resource Language Information Processing, 22(6). <https://doi.org/10.1145/3589131>

Wijayanti, R., & Arisal, A. (2021). Automatic Indonesian sentiment lexicon curation with sentiment valence tuning for social media sentiment analysis. ACM Transactions on Asian and Low-Resource Language Information Processing, 20(1). <https://doi.org/10.1145/3425632>

Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: A review. Artificial Intelligence Review, 53(6), 4335–4385. <https://doi.org/10.1007/S10462-019-09794-5>



Questions?

Thank You!