**Sentiment Analysis**

**Task Introduction**

This Realtime Dreamer project is to use real Ecommerce data to perform NLP for Vietnamese language and machine learning tasks in order to settle a real-world business challenge of a company in Vietnam. Some content of the online customer reviews does not reflect the ratings. For example, some with 5-rating reviews have negative comments and emotion such as positive, neutral and negative. Sentiment analysis for Vietnamese reviews is needed for company to better understand their customers’ needs in order to improve product and sales performance.

**Data Source**

| Created Method | Data provider | Format | Data size | Related Variables | Location |
| --- | --- | --- | --- | --- | --- |
| Clear positive and negative keyword search and eye scan to select reviews for emotion labeling | Yunhong He | Excel | 6018 rows, 4 columns | Index, Review Content, Rating, and emotion | [sample data](https://github.com/yunhonghe/realtime_dreamer/blob/main/data/processed/sentiment_analysis_reviews_label.xlsx) for train reviews.xlsx |
| manually labeled emotions for reviews | Suwasit’s Vietnam Team | Excel | 22771 rows, 19 columns | Review Content, 'Comment classified Type 1' | [sample data](https://github.com/yunhonghe/realtime_dreamer/blob/main/data/processed/Git_mockup_reviews_processed.xlsx) |
| Online Vietnamese emotion lexicon | [Kensuke Suzuki](https://app.slack.com/team/U01FHATGAAK) found online | Excel | 10627 rows, 12 columns | Vietnamese, Positive, Negative | [VnEmoLex](https://zenodo.org/record/801610#.YtnjGnbMJ7U) |

**Data Pre-processing**

| 1. Review label train data  There is imbalanced data in train reviews.xlsx. 8% of the labels are negative class while 87% of the labels are positive class. As the task mainly focuses on finding the reviews with 4 or 5 ratings but have negative emotions, F1 macro score is important for model evaluation. Oversampling negative class by random.choices function can make the size of negative class as same as that of positive class. |  | Graph 1: Imbalanced Emotion Classes |
| --- | --- | --- |

2. Train data provided by Suwasit’s Vietnamese team

| Reviews with labels of ‘Good’ or ‘Bad’ are labeled ‘positive’ or ‘negative’ correspondingly. Some reviews did not have labels due to mixing up both positive and negative emotions, and therefore are labeled with ‘neutral’. Furthermore, rows with null values in column 'Review Content', ‘Rating’ are removed. |  | | Class | Record count | | --- | --- | | Positive | 11271 | | Neutral | 10641 | | Negative | 859 |   Table 1: Emotion counts in dataset provided by Vietnamese team |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

3. [VnEmoLex](https://zenodo.org/record/801610#.YtnjGnbMJ7U) - Online Vietnamese emotion lexicon

| Vietnamese words with both labels of ‘Positive’ and ‘Negative’ are labeled ‘neutral’, words with only ‘Positive’ label or ‘Negative’ are labeled as ‘positive’ or ‘negative’ correspondingly. Moreover, rows with null values in column ‘Vietnamese’ are removed. |  |  | | Class | Record count | | --- | --- | | Positive | 2800 | | Neutral | 4673 | | Negative | 3149 |   Table 2: Emotion counts in [VnEmoLex](https://zenodo.org/record/801610#.YtnjGnbMJ7U) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

Above ‘positive’, ‘neutral’ and ‘negative’ labels are replaced with 2, 1, and 0 respectively for model training.

**Model Selection and Evaluation**

Hugging Face pre-trained BERT models and Supervised machine learning algorithms are trained and evaluated in the sentiment analysis of the reviews.

1. pre-trained BERT model are evaluated based on F1 scores (macro, micro and weighted), train and validation losses:

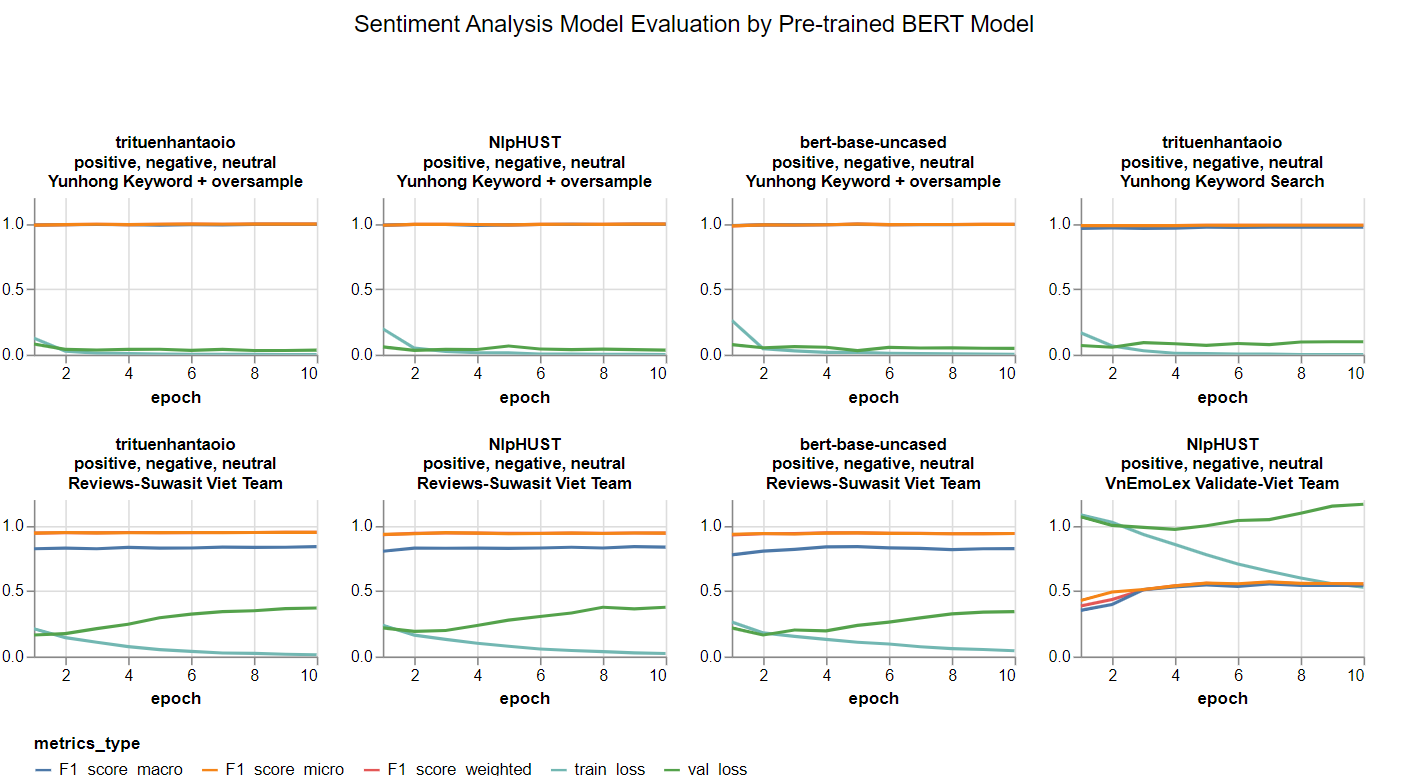
1)  [Hugging Face pre-trained BERT model trituenhantaoio/bert-base-vietnamese-uncased](https://huggingface.co/trituenhantaoio/bert-base-vietnamese-uncased).

2)  [Hugging Face pre-trained BERT model NlpHUST/vibert4news-base-cased](https://huggingface.co/NlpHUST/vibert4news-base-cased).

3)  [Hugging Face pre-trained BERT model NlpHUST/vibert4news-base-cased](https://huggingface.co/bert-base-uncased).

As shown in below Graph 2: Sentiment Analysis Model Evaluation by Pre-trained BERT Model,

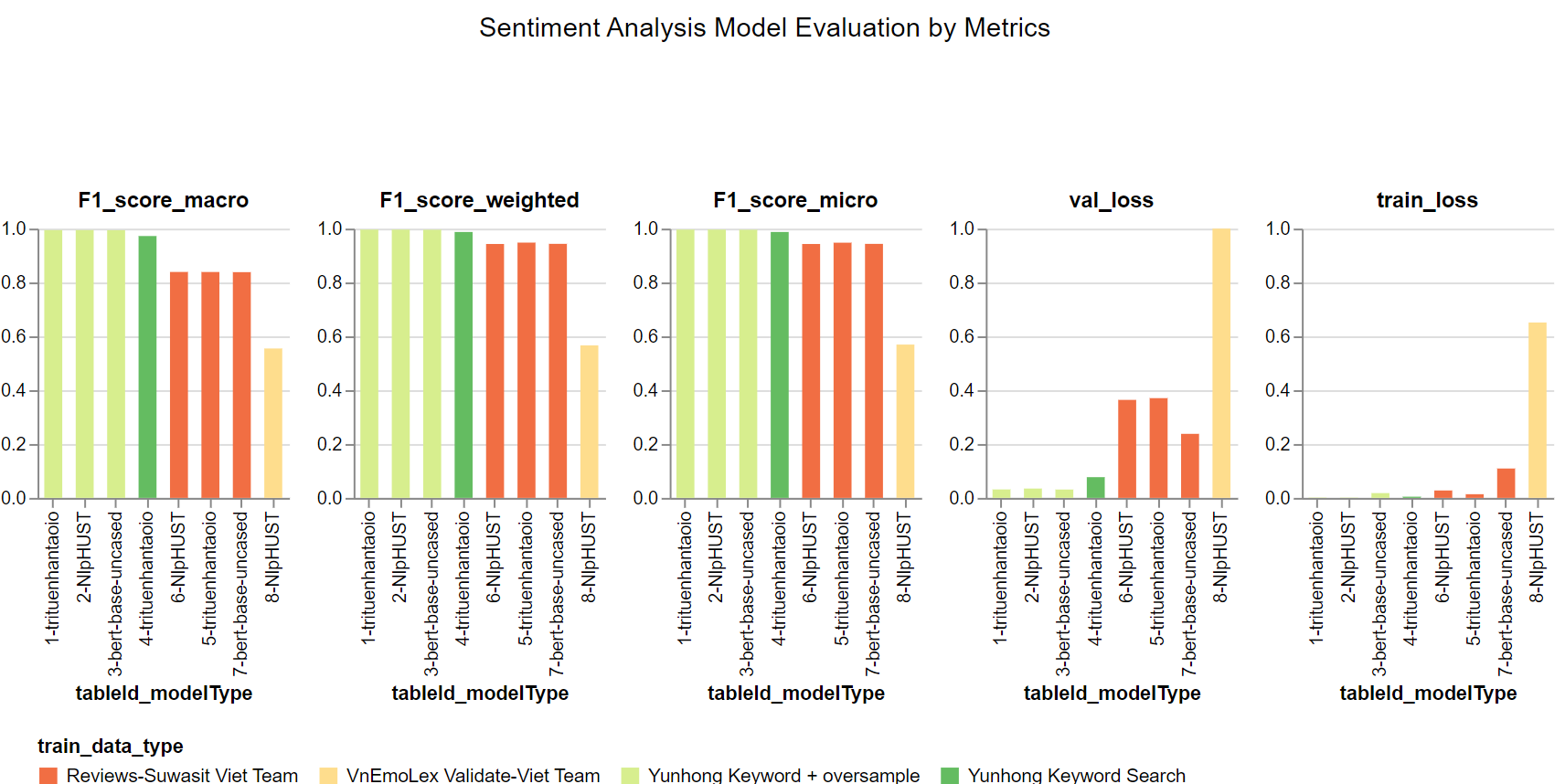
1. The combination of the method for labeling the emotions for online reviews by clear positive and negative keyword search and eye scan to select the reviews for emotion labeling and oversampling negative class significantly the classification performance of the BERT models. F1 scores (macro, micro, and weighted) are close to 1, train loss, and validation loss are close to zero across 10 epochs.
2. pre-trained BERT model trituenhantaoio/bert-base-vietnamese-uncased outperforms other pre-trained BERT models such as NlpHUST/vibert4news-base-cased and bert-base-uncased in terms of the stability of low train and validation losses across 10 epochs and the fast convergence after epoch 2.
3. Slightly higher validation losses across 10 epochs are shown in the 4th figure in the 1st row. This is because the only method used is keyword search and eye scan to label the reviews. Oversampling the minority negative class was not done.
4. Further Lower F1 scores and higher train and validation losses are shown in the figures in the 2nd row. The train data for the BERT models in these figures are not created by keyword search and eye scan. Especially the important F1 macro scores are lower than F1 micro and weighted scores in the same figures. Moreover, validation loss increases quite high as the epoch increases comparing with the those in the figures in the 1st row.
5. BERT model NIpHUST trained by the online Vietnamese emotion lexicon [VnEmoLex](https://zenodo.org/record/801610#.YtnjGnbMJ7U) dataset had very low F1 scores (macro, micro and weighted) ranging from 0.35 to 0.57, and the highest validation losses ranging 0.97 to 1.16 across 10 epochs. This is possibly because that 0.51% of words are both positive and negative, which confuses the BERT model. Moreover, the performance of the BERT model trained by only single Vietnamese words is most likely worse than those trained by the sentences which has context.



Graph 2: Sentiment Analysis Model Evaluation by Pre-trained BERT Model

As shown in below Graph 3: Sentiment Analysis Model Evaluation by Metrics, the similar findings are presented. The best epoch for the highest F1 macro score of each BERT model is computed, the performance of the BERT models with their F1 micro score, F1 weighted score, train and validation loses in their own best epoch are shown in Graph 3.

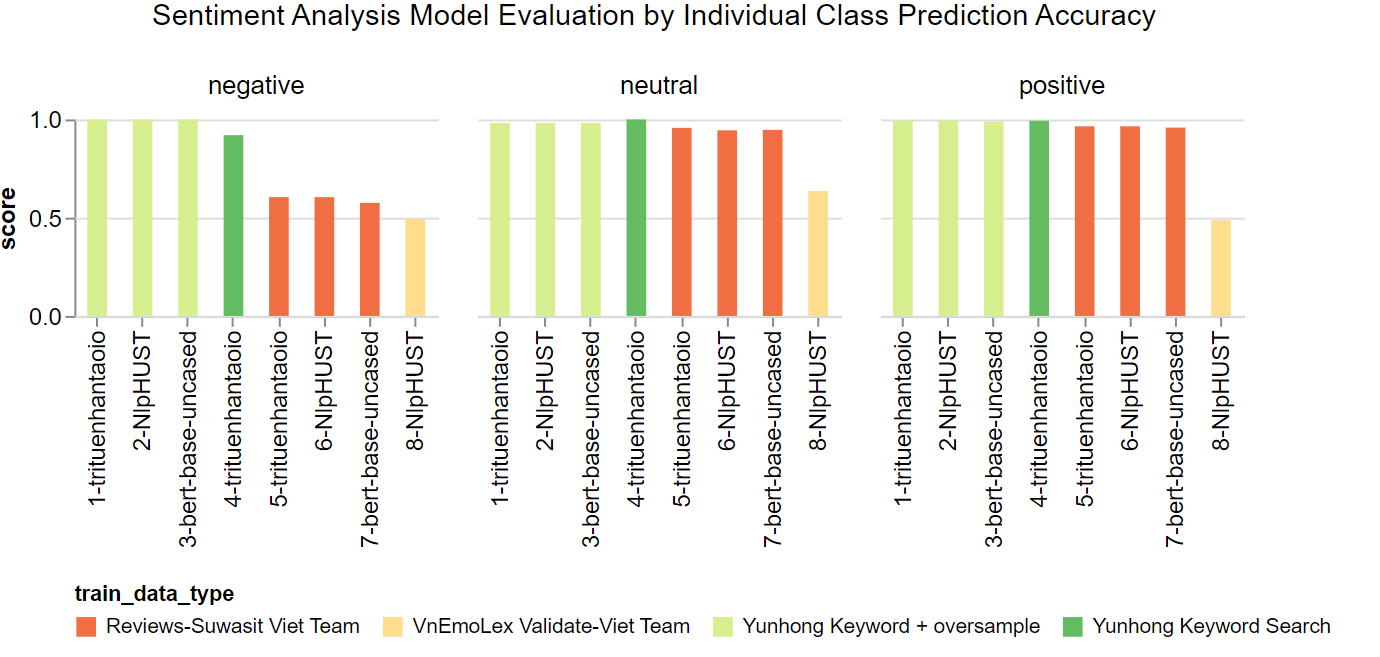
1. The BERT models trained by reviews label data created by clear positive and negative keyword search and eye scan, and oversampled negative class have the best classification performance, their F1 scores (macro, micro, and weighted) are close 1, train loss and validation loss are close to zero. outperforming those BERT models shown in the 2nd row in Graph 2, which were trained by online Vietnamese emotion lexicon [VnEmoLex](https://zenodo.org/record/801610#.YtnjGnbMJ7U) dataset or the dataset provided by Vietnamese team who manually labeled the reviews.
2. pre-trained BERT model trituenhantaoio/bert-base-vietnamese-uncased outperforms other pre-trained BERT models such as NlpHUST/vibert4news-base-cased and bert-base-uncased in terms of F1 scores (macro, micro and weighted), train and validation losses.
3. BERT model NIpHUST trained by the online Vietnamese emotion lexicon [VnEmoLex](https://zenodo.org/record/801610#.YtnjGnbMJ7U) dataset had the highest train loss of over 0.6 and validation loss of about 1, and lower than 0.6 of F1 scores (macro, micro and weighted).



Graph 3: Sentiment Analysis Model Evaluation by Metrics

Similar findings are also presented in Graph 4: Sentiment Analysis Model Evaluation by Individual Class Prediction Accuracy:

1. The BERT models trained by reviews label data created by clear positive and negative keyword search and eye scan, and oversampled negative class have the best classification performance, their accuracy scores for positive, neutral and negative classes are close to 1, outperforming those BERT models which were trained by online Vietnamese emotion lexicon [VnEmoLex](https://zenodo.org/record/801610#.YtnjGnbMJ7U) dataset or the dataset provided by Vietnamese team who manually labeled the reviews.
2. pre-trained BERT model trituenhantaoio/bert-base-vietnamese-uncased still has the best performance in terms of the accuracy of classification for positive, neutral and negative classes.
3. BERT model NIpHUST trained by the online Vietnamese emotion lexicon [VnEmoLex](https://zenodo.org/record/801610#.YtnjGnbMJ7U) dataset had the lowest accuracy scores ranging from 0.49 to 0.64 for positive, neutral and negative classes.



Graph 4: Sentiment Analysis Model Evaluation by Individual Class Prediction Accuracy

1. Supervised machine learning algorithms are evaluated based on F1 scores (macro, micro and weighted):

1)  [DummyClassifier(strategy="uniform")](https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html)

2)  [DummyClassifier(strategy="most\_frequent")](https://scikit-learn.org/stable/modules/generated/sklearn.dummy.DummyClassifier.html)

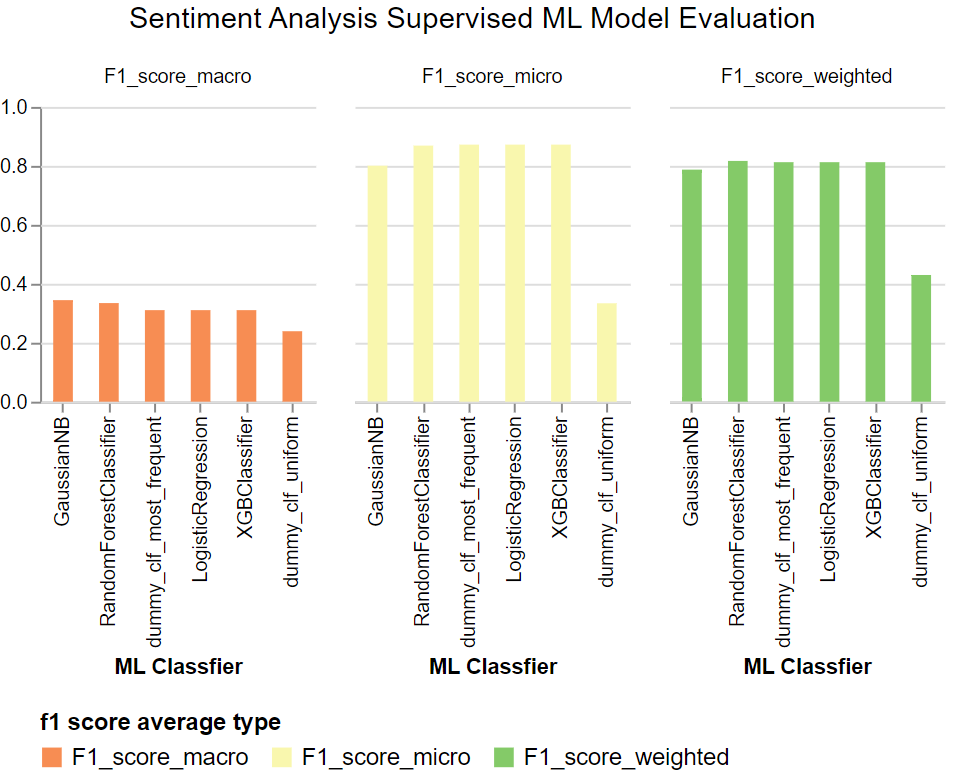
3)  [LogisticRegression(solver='lbfgs', multi\_class='auto')](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

4)  [RandomForestClassifier()](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)

5)  [GaussianNB()](https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html)

6)  [xgb.sklearn.XGBClassifier()](https://xgboost.readthedocs.io/en/stable/python/python_api.html)

As shown in below graph 5: Sentiment Analysis Supervised ML Model Evaluation, classification performance of Supervised ML classifiers is very poor in terms of less than 0.4 F1 macro score for all the models. F1 micro and weighted scores are also lower than those BERT models which are not trained by Vietnamese emotion lexicon [VnEmoLex](https://zenodo.org/record/801610#.YtnjGnbMJ7U) dataset.



Graph 5: Sentiment Analysis Supervised ML Classifier Model Evaluation

**Summary**

Hugging Face pre-trained BERT model trituenhantaoio/bert-base-vietnamese-uncased is chosen to conduct sentiment analysis for emotion classification of online customer reviews due to its best performance in terms of its stable and fast convergence of close to zero of train loss, validation loss, and close to 1 of F1 scores (macro, micro and weighted), and classification accuracy for positive, neutral and negative emotions across 10 epochs. The train data created by clear positive or negative keyword search and eye scan to select the reviews for emotion labeling can significantly improve the BERT model performance, and thus is used to train the BERT model in order to gain best performance of emotion classification for the company to better understand the needs of their customers in order to improve customer services and sales performance.

**Appendix**

1. The sentiment analysis pipeline is located at <https://github.com/yunhonghe/realtime_dreamer>. The <https://github.com/yunhonghe/realtime_dreamer/blob/main/sentiment_analysis.sh> is used to run the pipeline.
2. The zip file of [Sentiment\_Analysis\_BERT\_Model\_Evaluation.ipynb](https://github.com/yunhonghe/realtime_dreamer/blob/main/notebooks/Sentiment_Analysis_BERT_Model_Evaluation.ipynb) can preserve the model evaluation visualizations and is located at <https://github.com/yunhonghe/realtime_dreamer/blob/main/notebooks/Sentiment_Analysis_BERT_Model_Evaluation.7z>

**References**

1. trituenhantaoio/bert-base-vietnamese-uncased (Hugging Face)

<https://huggingface.co/trituenhantaoio/bert-base-vietnamese-uncased>

2. NlpHUST/vibert4news-base-cased (Hugging Face)

<https://huggingface.co/NlpHUST/vibert4news-base-cased>

3. bert-base-uncased (Hugging Face)

<https://huggingface.co/bert-base-uncased>