**VIETNAM GENRAL CONFEDERATION OF LABOR**

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**

****

**FINAL PROJECTS**

**Knowledge Discovery and Data Mining**

**Vietnamese Social Media Emotion Corpus**

*Instructor:* **LE CUNG TUONG**

*Moderators:***NGUYEN THANH KHANG -518H0072**

Class**: 18H50303**

Course**: 22**

HO CHI MINH CITY, 2021

**VIETNAM GENRAL CONFEDERATION OF LABOR**

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**

****

**FINAL PROJECTS**

**Knowledge Discovery and Data Mining**

**Vietnamese Social Media Emotion Corpus**

*Instructor:* **LE CUNG TUONG**

*Moderators:***NGUYEN THANH KHANG -518H0072**

Class**: 18H50303**

Course**: 22**

HO CHI MINH CITY, 2021

Contents

[ACKNOWLEDGMENTS. 4](#_Toc71473836)

[TEACHER'S CONFIRMATION AND REVIEW SECTION 6](#_Toc71473837)

[I. What is emotional analysis in natural language processing? 7](#_Toc71473838)

[II. Emotional text parser application 8](#_Toc71473839)

[III. Methods of solving emotional analysis problems. 9](#_Toc71473840)

[1. The method based on a dictionary of words to express emotions. 9](#_Toc71473841)

[2. Deep Learning Neural Network method. 9](#_Toc71473842)

[3. Rule-bases (Law-Based) and Corpus-bases Combination Methods. 9](#_Toc71473843)

[IV. Training diagram of Vietnamese text emotional analysis model (Sentiment Analysis Vietnamese - SAV). 10](#_Toc71473844)

[1. Coaching diagrams (training): 10](#_Toc71473845)

[2. Inspection (test) and commissioning (release) diagrams: 11](#_Toc71473846)

[V. Some notes on the tuning of the LSTM super parameter. 11](#_Toc71473847)

[VI. Practice creating Vietnamese Emotional Analysis (SAV) model. 14](#_Toc71473848)

[1. Collecting data: 14](#_Toc71473849)

[2. Emotion Recognition for Vietnamese Social Media Text: 15](#_Toc71473850)

[3. Applying the trained model in Task 2: 15](#_Toc71473851)

[VII. References 16](#_Toc71473852)

# ACKNOWLEDGMENTS.

This is the final article on our confidential information map, but we are still lacking, I hope you will give us more advice to continue to add knowledge and motivation for them. I am on the way to study. Thank you very much!

**THE SUBJECTS ARE COMPLETED**

**AT TON DUC THANG UNIVERSITY**

I hereby undertake that this is my own project / our project and under the guidance of Le CUNG TUONG . The research contents and results in this topic are truthful and have not been published in any form before. The data in the tables for analysis, comments and evaluation collected by the author from different sources are clearly stated in the references.

In addition, the project also uses a number of comments, assessments as well as data of other authors, other organizations and organizations with citations and origin notes.

If I find out there is any fraud I take full responsibility for the content of my project. Ton Duc Thang University is not related to the copyright and copyright violations caused by me in the implementation process (if any).

*TP. Ho Chi Minh City, 2021*

*Author.*

*Nguyen Thanh Khang.*

# 

# TEACHER'S CONFIRMATION AND REVIEW SECTION

**Verification of the instructor**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

City. Ho Chi Minh City, date, month ,year.

(sign and write full name)

**The teacher evaluation section marks**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

City. Ho Chi Minh City, date, month, year.

(sign and write full name)

# Introduction to Data Mining:

Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. The computer is responsible for finding the patterns by identifying the underlying rules and features in the data.

# Description of the problem and problem domain:

Emotional Analysis (Emotional Analysis) is aimed at discovering lasting mode, emotional color, Directing trust in the subjects or people.

Issues surrounding emotional analysis:

1. The source of emotion.
2. Emotional goal.
3. Types of emotions: Disgust (chán ghét), Enjoyment(thích thú), Anger (giận dữ), Surprise (ngạc nhiên), Sadness (buồn bã), Fear (sợ hãi),Other (khác).
4. About emotional level: positive, negative, neutral.
5. Emotional text: a sentence or a piece of text.

The emotional analysis problem belongs to the form of text semantic analysis problem. Therefore, we need to build a model to understand the meaning of a sentence or paragraph to decide which sentence or paragraph has the main emotional color.

Speaking from the perspective of machine learning (Machine Learning), emotional analysis is a problem of classifying emotions based on natural language text. The input of the problem is a sentence or piece of text, and the output is the probability values (scores) of the N emotional classes that we need to determine.

In types of emotional analysis problems are classified into problems of different difficulty as follows:

1. Simple: Analyze emotions (attitudes) in the text into 2 classes: positive and negative.
2. More complicated: Rank emotions (attitudes) in text from 1 to 6.

* Disgust (chán ghét),
* Enjoyment(thích thú),
* Anger (giận dữ),
* Surprise (ngạc nhiên),
* Sadness (buồn bã),
* Fear (sợ hãi),
* Other (khác).

1. Difficulty: Discovering the goals, sources of emotions (attitudes) or complex emotions (attitudes).

# Emotional text parser application

Emotional analysis in the text is used in a variety of issues such as corporate branding, product branding, customer relationship management, sociological opinion surveys, state analysis. human psychology...

We are living in the digital age, especially in recent years emerging with social networks, with millions of users around the world, with a huge amount of content information generated by users every day, with many form forms such as status lines, images, and videos. Social networks have the following characteristics: user-generated, personal information, so the quality of the content, or the correctness and authenticity, is relative; a newly created information can spread quickly to a large number of other users, compared to traditional information channels such as television, radio, newspapers, forums, blogs ...

This poses difficulties for large enterprises to solve the problem of corporate branding and product brand management in front of bad public opinion on social networks, both in terms of the source of information, both the amount of information to be processed. Not to mention that competitors in the marketplace take advantage of social networks to intentionally create unfavorable information for each other.

A specific example in Vietnam is the recent case of "a fly in the number one bottle" of Tan Hiep Phat enterprise, which adversely affected the image of Tan Hiep Phat and the consumption of energy drink products. of this business. In terms of law, Tan Hiep Phat is right but not skillful in handling customer relationships, causing discontent on social networks, it is a problem of relationship management with customers that businesses must solve. . Who knows if this unfavorable news about Tan Hiep Phat has been motivated by competitors? This requires a powerful support tool, which can only be solved by applying information technology, but not by any human force.

Learning from Tan Hiep Phat's experience, large Vietnamese enterprises have now ordered information technology enterprises to solve this problem. Current technology solutions are called "social media listening", which means that IT businesses buy real-time data from social networking companies to process-related information. to the business or the products, it does business, to detect and prevent the early spread of adverse information on social networks, to take the form of correcting feedback to its customers, and at the same time negotiate, to radically prevent people who create such content. The essence of this solution is to analyze the emotions of the social media status lines to filter out the negative information for processing.

# Methods of solving emotional analysis problems.

Currently, the emotional analysis problem has a number of solutions to solve the following:

## The method based on a dictionary of words to express emotions.

Accordingly, the prediction of emotions is based on finding individual emotional words, determining scores for positive words, determining scores for negative words, and then aggregating these scores. according to a definite metric that determines what emotional color the text is in. The drawback of this method is that the order of words is ignored and important information can be lost. The accuracy of the model depends on how well a dictionary of emotions is. However, it has the advantage of easy implementation, fast calculation cost, only effort in building a dictionary of emotions.

## Deep Learning Neural Network method.

In recent decades, with the rapid development of CPU and GPU processing speed and decreasing hardware costs, cloud infrastructure services increasingly develop, as the premise and opportunity. for deep learning method, Deep Learning Neural Network develops strongly. In particular, the problem of emotional analysis has been solved by the Recurrent Neural Network (RNN) learning model with a variant commonly used today is the Long Short Term Memory Neural Network (LSTMs), combined with the vector model. word (vector representations of words) Word2Vector with its Continuous Bag-of-Words (CBOW) architecture. This model gives more than 85% accuracy. The advantage of this method is that the input text can be a sentence or a paragraph. Implementing this model requires as much textual data as possible to create high-quality Word2Vector CBOW and large-labeled data to train, validate, and test the tissue. Supervise Learning LSTMs.

## Rule-bases (Law-Based) and Corpus-bases Combination Methods.

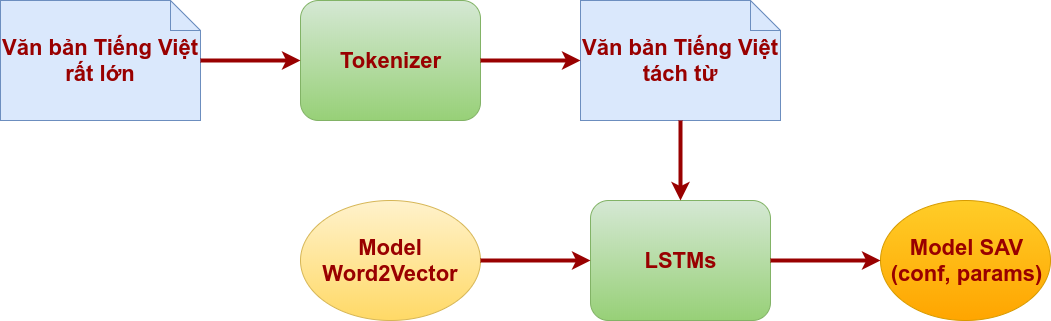
An example of this approach is the work of Richard Socher of Stanford University. You can refer to here: **http://nlp.stanford.edu/sentiment/**. This method combines the use of the Deep Learning Recursive Neural Network model with the expert knowledge system in natural language processing (XLNNTN) called Sentiment Treebank. A sentiment Tree is a parsing tree of a sentence, in which each node in the tree has a set of emotional weights in turn: very negative, negative, neutral, positive, and very positive. Accordingly, the largest label weight determines the global label of the node, as shown below. The model's accuracy when predicting emotions for a single sentence is 85.4%. The disadvantage of this method is that it only handles the input data well, in a single sentence.

Based on the above analysis, we decided to choose the deep learning method LSTMs combined with Word2Vector to solve the emotional analysis problem. This model appears to be close to the practical application requirements with the input text of any paragraph, which can be commented (comments) on social networks, reviews (reviews) on selling websites. restaurant, catering services, entertainment, tourism such as restaurants, restaurants, hotels, tourist attractions, movie theaters, movies, famous brands... The output is The emotional class falls into two categories: negative and positive.

# Description of implemented data mining techniques/methods

Training diagram of Vietnamese text emotional analysis model (Sentiment Analysis Vietnamese - SAV).

## Coaching diagrams (training):



As shown in the figure above, we see that the first part of the supervised learning model LSTMs is the file label, containing the word fragment (segment) processed text using the Tokenizer and model Word2Vector tools.

Vietnamese is a single-configuration type, with undefined word boundaries, naturally with white accounts, but solving the problem is related to the meaning of the word, so the segment is required to achieve the model. For higher accuracy, you can read the Vietnamese word isolation article to understand more.

The rest of the Word2Vector model is the result of the training process based on the Recurrent Neural Network (RNN) model to vectorize words, or in other words, put words into space vectors, you can read the article. How to create Word2Vector for Vietnamese to understand more.

As a result of the training, we get the critical network number of LSTMs saved to the file (params) along with the metamorphic LSTMs (conf) network configuration that we have set. This file will be download (download) to the LSTM network to test, operate (release) or be able to continue training (training) later.

## Inspection (test) and commissioning (release) diagrams:

We see at this test and operator stage, the model LSTMs upload important configuration files (conf) and the neural network's params file as a result of pre-training. there. At the same time, we must use the Word2Vector model as the vocabulary knowledge system.

During the inspection process, we include a data set including files containing fragments of text that were labeled with the Tokenizer tool earlier. The class results are recorded to compare with the original data header of the label, thereby giving us the exact results of the model.

If after too a test, the main of the model is being in a accept the level, we will use the results of this model into execution. When the model is inserted into the operating system, the input is the undefined label and the purpose of the generated emotional model analysis is to attach the data head to this label, the host for the purpose class. Data. The first predicted result is accepted with a reliable error definition.

# Some notes on the tuning of the LSTM super parameter.

In the article, we have not analyzed the content data of the short-term memory network (LSTM) neural network, for other purposes the article is not too long, we may present LSTM neural networks in one article. Instead, let us quickly dive into how to effectively use LSTMs, the possible tool in this case for solving emotional analytical text.

In the field of Deep Learning in general and LSTM networks in particular, it is difficult to adjust metadata when creating a model network, which is quite difficult for all those with long experience, not people. new or beginner to learn about Deep Learning. Selecting each network is a complex model or simple but the metadata number will also be more or less corresponding.

To put it plainly, we imagine adjusting the deep learning metadata as if we were adjusting the sound of karaoke amplifiers, with more than 20 different buttons and buttons. Inexperienced users have to try wrong many times to get the sound satisfied, when users with long experience, just try 3-4 times to succeed.

One thing to keep in mind is that the cost of time to try and know yourself wrong when training deep learning networks is measured in hours, by days, for example, the training time for model SAV in this post ends at 10:20 a.m. for the bad time test have a think all for big-time on the time and the power off.

In large laboratories, training time can be reduced by using more machines to process songs, using GPUs to increase processing power, but requiring a pre-investment. big. That is why we try to write this section to highlight some of the implementation experiences that scientists have put together to limit the trial and error problems to the lowest possible level. But when you embark on an experiment, you will find that all these things are poignant.

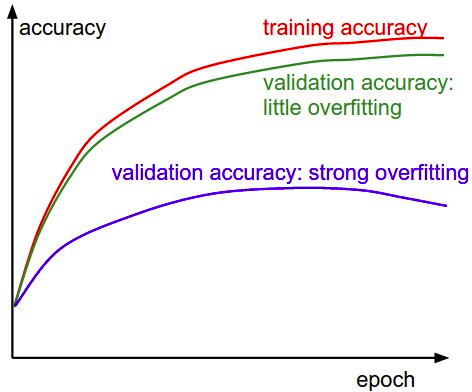
For the sake of short writing, we will only present considerations when tuning the metadata for network LSTMs. When you have more experience, it is automatically found that this adjustment is also normal, the problem is that you have to understand the meaning of each parameter that plays a role in the deep learning network that you use.

1. First is Regulation methods, which are the methods of fine-tuning or correcting geometries, to solve overfitting problems. With LSTM network, scientists recommend using 2 methods named L1 and L2, in the SAV model we choose to use L2
2. Test data set independent of training data. To evaluate the customer more objectively about the accuracy of the model.
3. The larger the network is, the stronger it is, but it is also prone to overfit. Never try to train a model with some parameter (critical number) with only 10,000 sample data. We need to pay attention to the rule:

**parameters > example = trouble**

That is, when the number of parameters is greater than the number of samples, we will have problems with overfitting.

1. More coaches, almost always better, resist overfitting as outlined above.
2. Retrain many times (epochs). With the same training data definition, the model will be relearned over that data multiple times to improve the model's accuracy over each learning time. These hyperparametric epochs usually range from 1-5 times. Of course, theoretically, the higher the era, the better, but in return, it is the cost of long-term training. It's like "chew it well not long, plow the rice well". The model will learn more from the data with more training.
3. There is always a file review to evaluate across each era to know when it needs to end (possibly ending soon).



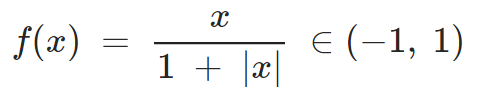
In the training process, at the end of a training session of the model (epoch), we should have an evaluation set to evaluate the accuracy of the model through each learning (epoch). If we observe increasing accuracy with each study, we can rest assured that the model is not overfitting. On the contrary, when it comes to the second learning (n + 1), we see that the model's accuracy is reduced from the previous n times and continues to decrease at the second (n + 2), we should end too early the learning process to refine the metadata, then retrain, to avoid wasting time in vain.

1. The learning rate is a very important parameter. It affects the error of the model throughout the learning process.



As the picture above shows us four possible scenarios when setting the learning rate value during model training. Looking at the shape of the model's error graph over the times we can see that we set the initial learning rate as low, high or too high compared to the best value, in order to stop working in time. retraining to adjust the learning rate.

1. For LSTM, use softsign (not softmax) activation function. Faster and less saturated (~ 0 gradient).



1. Updaters: RMSProp, AdaGrad or Momentum (Nesterovs) is used to use for supermet ratio to be used in the too training. These three algorithms are good choices for LSTMs network compared to other algorithms. For model SAV, Updater is selected as RMSProp.
2. Parameter initialization (key number) with Xavier.
3. Data normalization (data normalization) with the use of the error MCXENT function (MCXENT loss function) and the softmax activation function for computational regression (regression). Here, softmax is used to specify the calculation in the entire partition system, not to be confused with the softsign activation function located inside the LSTMs in Section 8 above.

Above are some notes when using Internet LSTMs, including suggestions in solving emotional analysis problems in Vietnamese (Vietnamese Emotional Analysis - SAV). The following is a more detailed description of how to use LSTMs in the experimental SAV model.

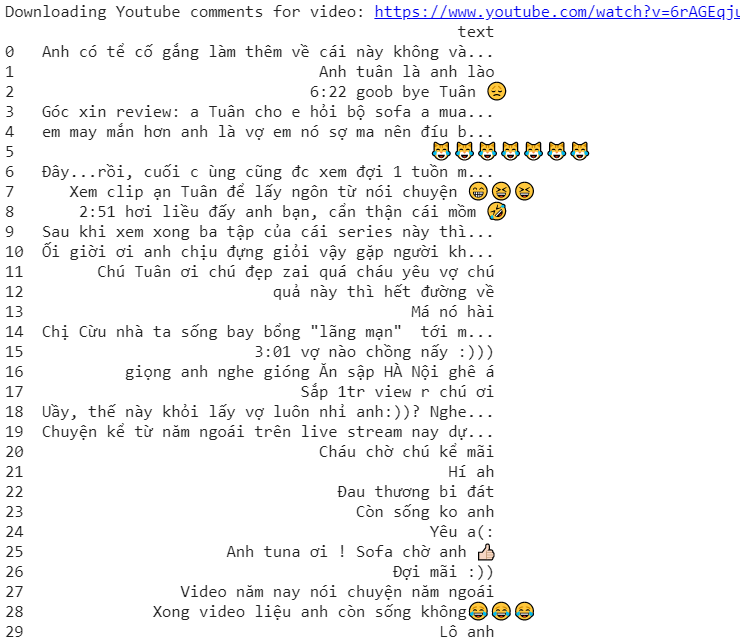
# Practice creating Vietnamese Emotional Analysis (SAV) model.

## Collecting data:

1. Choosing three Vietnamese videos on YouTube with at least 100 comments. Crawling the first 100 comments of each videos.

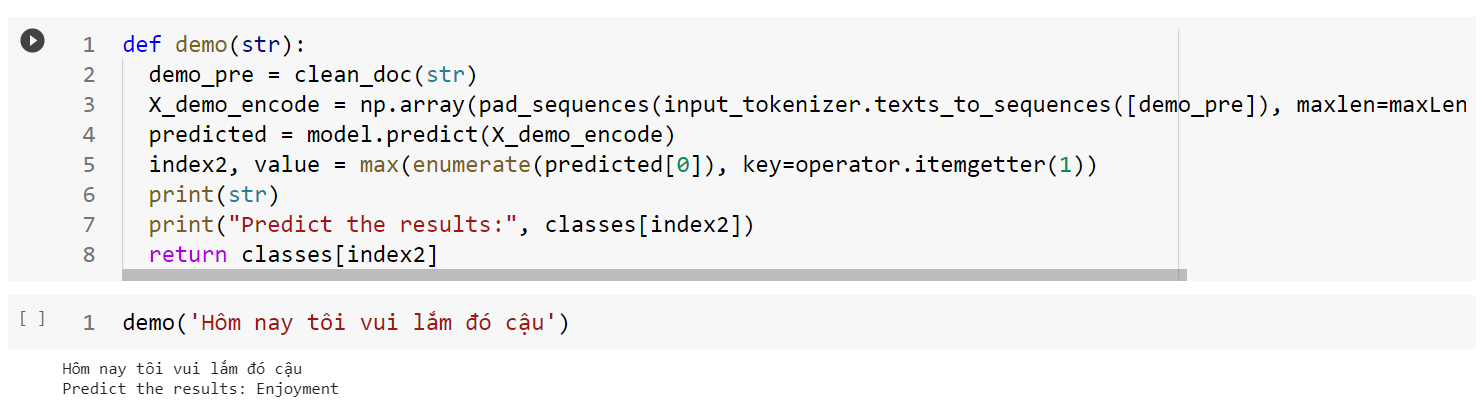
Please find the Crawl Data Example.ipynb for an example of crawling.

1. Label these comments by yourself in seven classes: Disgust (chán ghét), Enjoyment(thích thú), Anger (giận dữ), Surprise (ngạc nhiên), Sadness (buồn bã)Fear (sợ hãi),Other (khác).



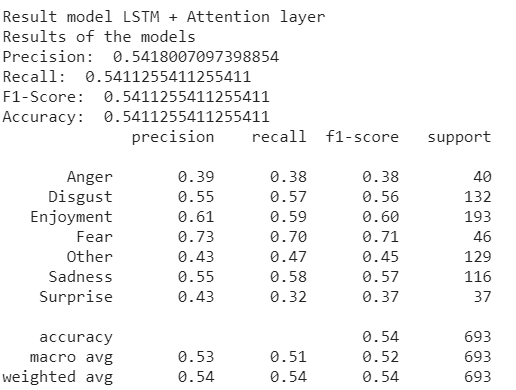
## Emotion Recognition for Vietnamese Social Media Text:

1. Using training and validation sets of UIT-VSMEC (UIT-VSMEC.zip) for training your Emotion Recognition model (classification model).



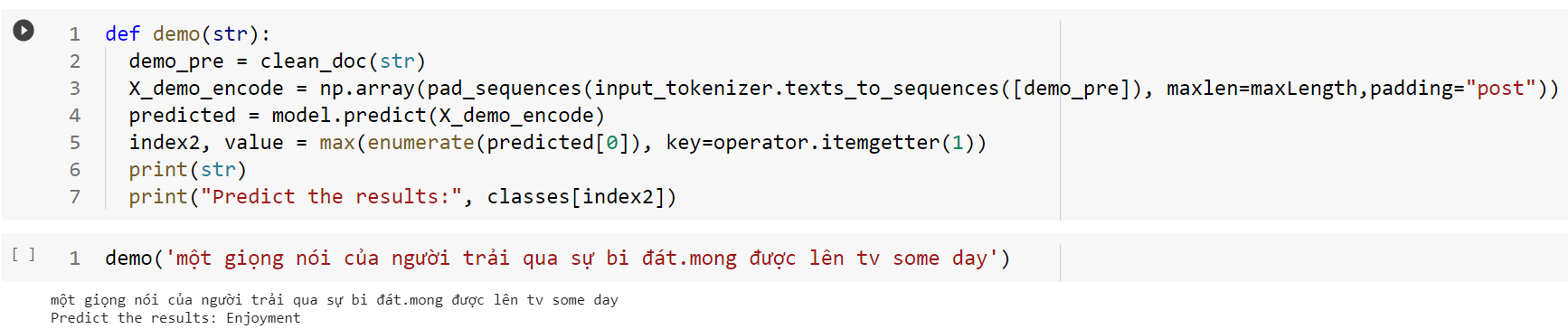
Note: you can choose any technique in preprocessing, any classifier that you want.

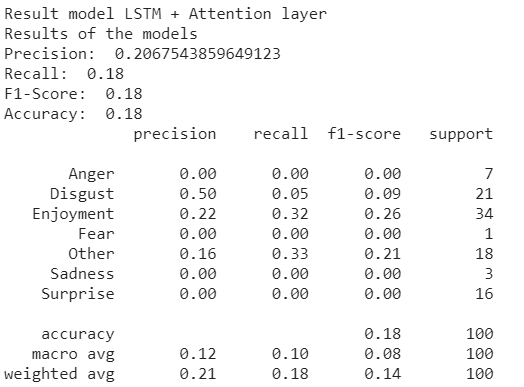
1. Report the performance metrics (Accuracy, F1-score...) for the test set in UIT-VSMEC dataset.



## Applying the trained model in Task 2:

Using three datasets in Task 1. Then,report the performance metrics (Accuracy, F1-score...) for these datasets.





# References

1. [SCV (streetcodevn.com)](https://streetcodevn.com/blog/sav)
2. [undertheseanlp/sentiment: Vietnamese Sentiment Analysis (github.com)](https://github.com/undertheseanlp/sentiment)
3. [concalahan/Sentiment-Analysis-Vietnamese: A simple model with 1 CNN layer and 2 LSTM layer for sentiment analysis problem. (github.com)](https://github.com/concalahan/Sentiment-Analysis-Vietnamese)
4. [KhiemLe99/Sentiment-Analysis-SoMeT-2020: A Novel Framework for Vietnamese Sentiment Classification (github.com)](https://github.com/KhiemLe99/Sentiment-Analysis-SoMeT-2020)
5. [(PDF) Emotion Recognition for Vietnamese Social Media Text (researchgate.net)](https://www.researchgate.net/publication/337438817_Emotion_Recognition_for_Vietnamese_Social_Media_Text)
6. [Vietnamese Social Media Emotion Corpus (UIT-VSMEC) dataset - NLP Database (metatext.io)](https://metatext.io/datasets/vietnamese-social-media-emotion-corpus-(uit-vsmec))