VIETNAM LABOR UNION GENERAL

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**FINAL PROJECT**

**Knowledge Discovery and Data Mining**

*Instructor:* **LÊ CUNG TƯỞNG**

*Moderators:* **NGUYỄN THANH TÙNG – 51800649**

**ĐÀO HOÀNG GIANG -518H0088**

Class**: 18H50301**

Course**: 22**

**HO CHI MINH CITY, 2021**

VIETNAM LABOR UNION GENERAL

**TON DUC THANG UNIVERSITY**

**FACULTY OF INFORMATION TECHNOLOGY**



**FINAL PROJECTS**

**Collecting data and Emotion recognition for Vietnamese Social Media Text**

*Instructor:* **LÊ CUNG TƯỞNG**

*Moderators:* **NGUYỄN THANH TÙNG – 51800649**

**ĐÀO HOÀNG GIANG -518H0088**

Class**:18H50301**

Course**: 22**

**HO CHI MINH CITY, 2021**

# ACKNOWLEDGMENTS.

I would like to thank you teacher for guiding me during the time project time.

**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.*

*(sign and write full name)*

*Nguyen Thanh Tung.*

*Dao Hoang Giang*

# 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)

# SUMMARY

Abstraction has its own group of refactoring techniques, mainly involving moving functionality along the class inheritance hierarchy, creating new classes and interfaces, replacing inheritance with authorization, and vice versa. again.

# TABLE OF CONTENTS

Contents

[ACKNOWLEDGMENTS. 3](#_Toc71800110)

[TEACHER'S CONFIRMATION AND REVIEW SECTION 4](#_Toc71800111)

[SUMMARY 5](#_Toc71800112)

[TABLE OF CONTENTS 5](#_Toc71800113)

[LIST OF SIGNS AND ABBREVIATIONS 6](#_Toc71800114)

[SIGNAL 6](#_Toc71800115)

[ABBREVIATIONS 6](#_Toc71800116)

[LIST OF TABLES, FIGURES, GRAPHICS 6](#_Toc71800117)

[LIST OF FIGURES 6](#_Toc71800118)

[LIST OF TABLES 6](#_Toc71800119)

[**CHAPTER 1: OVERVIEW.** 6](#_Toc71800120)

[**CHAPTER 2: THEORETICAL BASIS / EXPERIMENTAL RESEARCH.** 7](#_Toc71800121)

[2 .1 Collecting data 7](#_Toc71800122)

[2.2 Emotion Recognition for Vietnamese Social Media Text 9](#_Toc71800123)

[2.3 Applying the trained model in 2.2 to three datasets in 2.1. Then, report the performance metrics (Accuracy, F1-score...) for these datasets. 9](#_Toc71800124)

[**CHAPTER 3: RESULT EVALUATION**. 9](#_Toc71800125)

[REFERENCING DOCUMENTS 9](#_Toc71800126)

# LIST OF SIGNS AND ABBREVIATIONS

## SIGNAL

## ABBREVIATIONS

# LIST OF TABLES, FIGURES, GRAPHICS

## LIST OF FIGURES

## LIST OF TABLES

# CHAPTER 1: OVERVIEW.

**Knowledge discovery and data mining (KDD**) is an interdisciplinary field that focuses on methodologies to draw useful knowledge from data. The rapid growth of online data due to the Internet and the widespread use of databases has created a great demand for KDD methodologies. Challenges draw knowledge from research-based data in statistics, databases, pattern recognition, machine learning, data visualization, optimization and high-performance computation, to provide solutions for advanced web discovery and business intelligence.

IBM Research has been at the forefront of this exciting new field from its very inception. For more than a quarter of a century, an active program of statistical research has explored a wide range of theoretical and practical problems. Benoit Mandelbrot's pioneering work on fractal and long-range statistical models has had a significant impact on many disciplines, including hydrology, finance, media networks, and computer system analysis. Time-dependent data analysis and non-standard distributions is another influential area of ​​IBM statistical research. One example is the L moment distribution theory that has resulted in innovative statistical methods for characterizing and estimating distributions, especially data with large concentrations in finance, risk management, and IT system monitoring. A leader in knowledge discovery and data mining (KDD) was established in the 1990s by Rakesh Agrawal introduced association rule mining. Other major contributions by IBM in KDD include exploiting excessive information flow throughput using lightweight data analysis techniques, high-performance mining techniques in parallel execution environments, and industry pioneering. data mining to protect privacy.

With the explosive growth of online data and the expansion of IBM's service and consulting offerings, data-based solutions are becoming increasingly important. Accordingly, the development of methodology for business intelligence, as well as IT systems and business process monitoring, has become the focus of KDD and statistical research at IBM. In these areas, monitoring data that has been collected over time is used to make processes more efficient, more efficient, predictable, and profitable. Challenging aspects include processing time-dependent big data with different characteristics, creating accurate and realistic forecasting methods, and developing analytics that match decision-making. business plan. Two specific issues that IBM Research is currently addressing, such as customer targeting and business data forecasting.

# CHAPTER 2: THEORETICAL BASIS / EXPERIMENTAL RESEARCH.

## 2 .1 Collecting data

**How to Scrape YouTube Comments with Python**

So first, don’t forget to install a ChromeDriver from right here. You should also have Google Chrome installed. Now that this is done, let’s import the libraries we will need:

|  |
| --- |
| **import** time  **from** selenium.webdriver **import** Chrome  **from** selenium.webdriver.common.by **import** By  **from** selenium.webdriver.common.keys **import** Keys  **from** selenium.webdriver.support.ui **import** WebDriverWait  **from** selenium.webdriver.support **import** expected\_conditions **as** EC  **from** selenium **import** webdriver |

All the other modules are needed because YouTube comments are dynamically loaded, which means that they are only visible when you scroll down the page. So we want a loop that will:

* Scroll down
* Wait for comments to appear
* Scrape the comments
* Repeat for whatever range we want.

Here is the loop that does just that.

|  |
| --- |
| **def** CrawlingComment(url):  data=[]  options = webdriver.ChromeOptions()  options.binary\_location = r"C:**\P**rogram Files**\G**oogle**\C**hrome Beta**\A**pplication**\c**hrome.exe"  **with** Chrome(executable\_path=r'C:**\P**rogram Files**\c**hromedriver.exe',chrome\_options=options) **as** driver:  wait = WebDriverWait(driver,15)  driver.get(url)    **for** item **in** range(200):  wait.until(EC.visibility\_of\_element\_located((By.TAG\_NAME, "body"))).send\_keys(Keys.END)  time.sleep(2)    **for** comment **in** wait.until(EC.presence\_of\_all\_elements\_located((By.CSS\_SELECTOR, "#content-text"))):  data.append(comment.text)  **return** data |

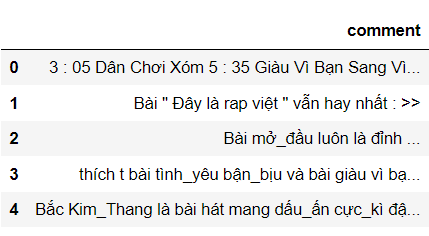
**So here is how it works:**

1. Access the URL you want with the driver.get function.
2. Scroll down and wait until everything is visible with wait.until and EC.visibility\_of\_element\_located.
3. Scrape the comments by finding all the #content-text elements (which is what we want, as you can see below) in the current viewed page.
4. 2.1.1 Choosing three Vietnamese videos on YouTube with at least 100 comments. Crawling the first 100 comments of each videos.
5. Append the comments to the data list.

To scrape comments from another video, all you need to do is change the URL! It’s that easy.

|  |
| --- |
| **import** pandas **as** pd  df = pd.DataFrame(data, columns=['comment'])  df.head() |

**After retrieving the data, we get the following results:**

****

**When there is comment data on YouTube, we proceed to assign labels to each data line. This step you can do it manually if you are free. Here we use LSTM to perform automatic labeling**

Self-label these comments in seven classes: Disgusted (disgusted), Enjoyed (amused), Anger (angry), Surprise (surprised), Sadness (sad), Fear (fear), Other).

**And get results:**

****

## 2.2 Emotion Recognition for Vietnamese Social Media Text

Emotion recognition or emotion prediction is a higher approach or a special case of sentiment analysis. In this task, the result is not produced in terms of either polarity: positive or negative or in the form of rating (from 1 to 5) but of a more detailed level of analysis in which the results are depicted in more expressions like sadness, enjoyment, anger, disgust, fear and surprise. Emotion recognition plays a critical role in measuring brand value of a product by recognizing specific emotions of customers’ comments. In this study, we have achieved two targets. First and foremost, we built a standard Vietnamese Social Media Emotion Corpus (UIT-VSMEC) with exactly 6,927 emotion-annotated sentences, contributing to emotion recognition research in Vietnamese which is a low-resource language in natural language processing (NLP). Secondly, we assessed and measured machine learning and deep neural network models on our UIT-VSMEC corpus.

**Data preprocessing**

**Word Separator (Tokenizer)** Vietnamese does not use morphemes to create word meanings (in English, cup->cups, 's' is the plural form). So in Vietnamese, words are not changed. To create different shades of meaning, Vietnamese depends on word order.

|  |
| --- |
| **def** tokenizer(text):  token = ViTokenizer.tokenize(text)  **return** token |

**Common word removal**

We can also remove commonly occurring words from our text data First, let’s check the 10 most frequently occurring words in our text data.

**Rare word removal**

This is very intuitive, as some of the words that are very unique in nature like names, brands, product names, and some of the noise characters, such as html left outs, also need to be removed for different NLP tasks. We also use a length of the words as a criterion for removing words with very a short length or a very long length

**Spelling Correction**

Social media data always messy data and it has spelling mistakes. Hence, spelling correction is a useful pre-processing step because this will help us to avoid multiple words. Example, “text” and “txt” will be treated as different words even if they are used in the same sense. This can be done by text blob library

**Emoji removal**

Emojis are part of our life. Social media text has a lot of emojis. We need to remove the same in our text analysis

|  |
| --- |
| **def** deleteIcon(text):  text = text.lower()  s = ''  pattern = r"[a-zA-ZaăâbcdđeêghiklmnoôơpqrstuưvxyàằầbcdđèềghìklmnòồờpqrstùừvxỳáắấbcdđéếghíklmnóốớpqrstúứvxýảẳẩbcdđẻểghỉklmnỏổởpqrstủửvxỷạặậbcdđẹệghịklmnọộợpqrstụựvxỵãẵẫbcdđẽễghĩklmnõỗỡpqrstũữvxỹAĂÂBCDĐEÊGHIKLMNOÔƠPQRSTUƯVXYÀẰẦBCDĐÈỀGHÌKLMNÒỒỜPQRSTÙỪVXỲÁẮẤBCDĐÉẾGHÍKLMNÓỐỚPQRSTÚỨVXÝẠẶẬBCDĐẸỆGHỊKLMNỌỘỢPQRSTỤỰVXỴẢẲẨBCDĐẺỂGHỈKLMNỎỔỞPQRSTỦỬVXỶÃẴẪBCDĐẼỄGHĨKLMNÕỖỠPQRSTŨỮVXỸ,.\_]"  **for** char **in** text:  **if** char !=' ':  **if** len(re.findall(pattern, char)) != 0:  s+=char  **elif** char == '\_':  s+=char  **else**:  s+=char  s = re.sub('**\\**s+',' ',s)  **return** s.strip() |

**Emoticon’s removal**

In previous steps, we have removed emoji. Now, going to remove emoticons. What is the difference between emoji and emoticons? :-) is an emoticon and 😜 → emoji?

**Converting Emoji and Emoticons to words**

In sentiment analysis, emojis and emoticons express an emotion. Hence, removing them might not be a good solution.

**Executing with train and test files**

In this part we use the data set provided

Data consists of 4 files:

* Stopwords.txt
* Test\_nor\_811.csv
* Train\_nor\_811.csv
* valid\_nor\_811.csv

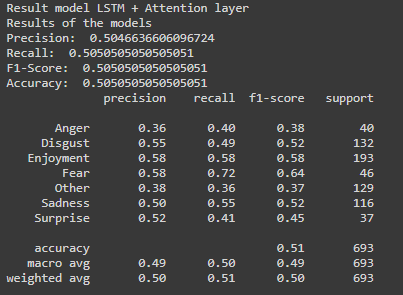
Applies to LSTM. mode

Short-term memory [1] (English: Long-term short-term memory, abbreviated LSTM) is an artificial recurrent neural network (RNN) used in the field of learning. Unlike the communication neural network (FNN) standard, LSTM contains connection responses. The network does not only process single point data (such as images), but also entire sequence data (as limited as speech or video). For example, LSTM can be applied to handwriting formats, speech recognition, and connected, non-segmented anomaly detection in information networks or IDS (import goods to the delivery system).

A unit LSTM information usually consists of a cell (cell), an input (input) port, an output (output) port, and a forget (forgot port). The cell writes values ​​at any interval and the three gates regulate the flow of information in/out the cell.

|  |
| --- |
| filter\_nums = 256 *# best 128*  **def** build\_model():  inputs = Input(shape=(maxLength, ), dtype='float64', name='inputs')  embedding\_layer = Embedding(input\_vocab\_size,EMBEDDING\_DIM,weights=[embedding\_matrix], input\_length=maxLength, trainable=True,name = 'word\_emb')(inputs)  embedding\_layer = SpatialDropout1D(0.75)(embedding\_layer)      lstm\_feature1 = CuDNNLSTM(filter\_nums, return\_sequences=True)(embedding\_layer)    att1 = AttentionWithContext()(lstm\_feature1)  att1 = Addition()(att1)    fc1 = Dropout(0.5)(Dense(256, name = 'dense\_1')(att1))  output1 = Dense(len(classes),name="output1", activation='softmax')(fc1)      *# define optimizer*    model = Model(inputs=inputs, outputs=output1)  tensorBoardCallback = TensorBoard(log\_dir='./logs', write\_graph=True)    model.compile(loss = 'categorical\_crossentropy', optimizer='adam', metrics=['accuracy'])      *# history = model.fit(X\_train\_encode, np.array(y\_train\_encode), validation\_data = (X\_val\_encode,np.array(y\_val\_encode)) , batch\_size=50, epochs=100,callbacks=[tensorBoardCallback])*  **return** model    model = build\_model() |

**Report the performance metrics (Accuracy, F1-score...)**

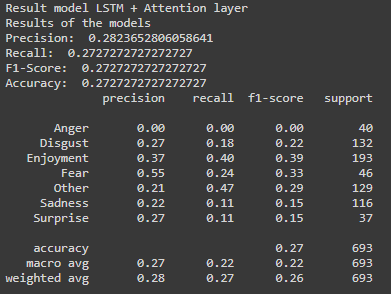
****

**Enter the demo program into 1 sentence**

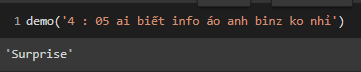
****

## 2.3 Applying the trained model in 2.2 to three datasets in 2.1. Then, report the performance metrics (Accuracy, F1-score...) for these datasets.

**Report the performance metrics (Accuracy, F1-score...)**

****

**Enter the demo program into 1 sentence**

****

# REFERENCING DOCUMENTS

**Vietnamese:**

<https://news.timviec.com.vn/kham-pha-data-mining-la-gi-60492.html>

**English:**

1.Bhowmick, P.K., Basu, A., Mitra, P.: An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Tex. In: Proceed-ings of the workshop on Human Judgements in Computational Linguistics. pp.

58–65. COLING 2008, Manchester, United Kingdom (2008)

2. company, J.S.: The habit of using social networks of Vietnamese people 2018.

brands Vietnam, Ho Chi Minh City, Vietnam (2018)

3. Ekman, P.: In: Facial expression and emotion. vol. 48, pp. 384–392. American Psychologist (1993)

4. Ekman, P.: In: Emotions revealed: Recognizing faces and feelings to improve com-munication and emotional life. p. 2007. Macmillan (2012)

5. Ekman, P., Ekman, E., Lama, D.: In: The Ekman’s’ Atlas of Emotion (2018)

6. Kim, Y.: Convolutional Neural Networks for Sentence Classifications. In: Proceed-ings of the 2014 Conference on Empirical Methods in Natural Language Process-ing (EMNLP). pp. 1746–1751. Association for Computational Linguistics, Doha, Qatar (2014)

7. Kravchenko, S., Mohammad, S.: Using Hashtags to Capture Fine Emotion Cate-goriest from Tweets. In: Computational Intelligence. pp. 301–326 (2015

8. Klinger, R., Clerc, O.D., Mohammad, S.M., Balladur, A.: IEST:WASSA-2018 Implicit Emotions Shared Task. pp. 31–42. 2017 AFNLP, Brussels, Belgium (2018)

9. Kratz Wald, B., Ilic, S., Kraus, M., S. Feuer Riegel, H.P.: Decision support with text-based emotion recognition: Deep learning for affective computing. pp. 24 – 35. Decision Support Systems (2018)

10. Mohammad, S., Bravo-Marquez, F.: Emotion Intensities in Tweets. In: Proceedings of the Sixth Joint Conference on Lexical and Computational Semantics (\*SEM). pp. 65–77. Association for Computational Linguistics, Vancouver, Canada (2017)

11. Mohammad, S.M.: #Emotional Tweets. In: First Joint Conference on Lexical and Computational Semantics (\*SEM). pp. 246–255. Association for Computational Linguistics, Montreal, Canada (2012)

12. Mohammad, S.M., Bravo-Marquez, F., Salameh, M., Kravchenko, S.: SemEval-2018 task 1: Affect in tweets. pp. 1–17. Proceedings of International Workshop on Semantic Evaluation, New Orleans, Louisiana (2018)

13. Mohammad, S.M., Xiaodan, Z., Kiritchenko, S., Martin, J.: Sentiment, emotion, purpose, and style in electoral tweets. pp. 480–499. Information Processing and Management: An International Journal (2015)

14. Nguyen: Vietnam has the 7th largest number of Facebook users in the world. DanTri newspaper (2018)

15. Nguyen, H.T.M., Nguyen, H.V., Ngo, Q.T., Vu, L.X., Tran, V.M., Ngo, B.X., Le, C.A.: VLSP Shared Task: Sentiment Analysis. In: Journal of Computer Science and Cybernetics. pp. 295–310 (2018)

16. Nguyen, K.V., Nguyen, V.D., Nguyen, P., Truong, T., Nguyen, N.L.T.: UIT-VSFC: Vietnamese Students’ Feedback Corpus for Sentiment Analysis. In: 2018 10th In-ternational Conference on Knowledge and Systems Engineering (KSE). pp. 19–24. IEEE, Ho Chi Minh City, Vietnam (2018)

17. Nguyen, P.X.V., Truong, T.T.H., Nguyen, K.V., Nguyen, N.L.T.: Deep Learning versus Traditional Classifiers on Vietnamese Students’ Feedback Corpus. In: 2018 5th NAFOSTED Conference on Information and Computer Science (NICS). pp.75–80. Ho Chi Minh City, Vietnam (2018)