WEEK 4

Scikit-learn Classification

Material:

1.Supervised Learning: [1. Supervised learning — scikit-learn 1.5.2 documentation](https://scikit-learn.org/stable/supervised_learning.html?authuser=0#supervised-learning)

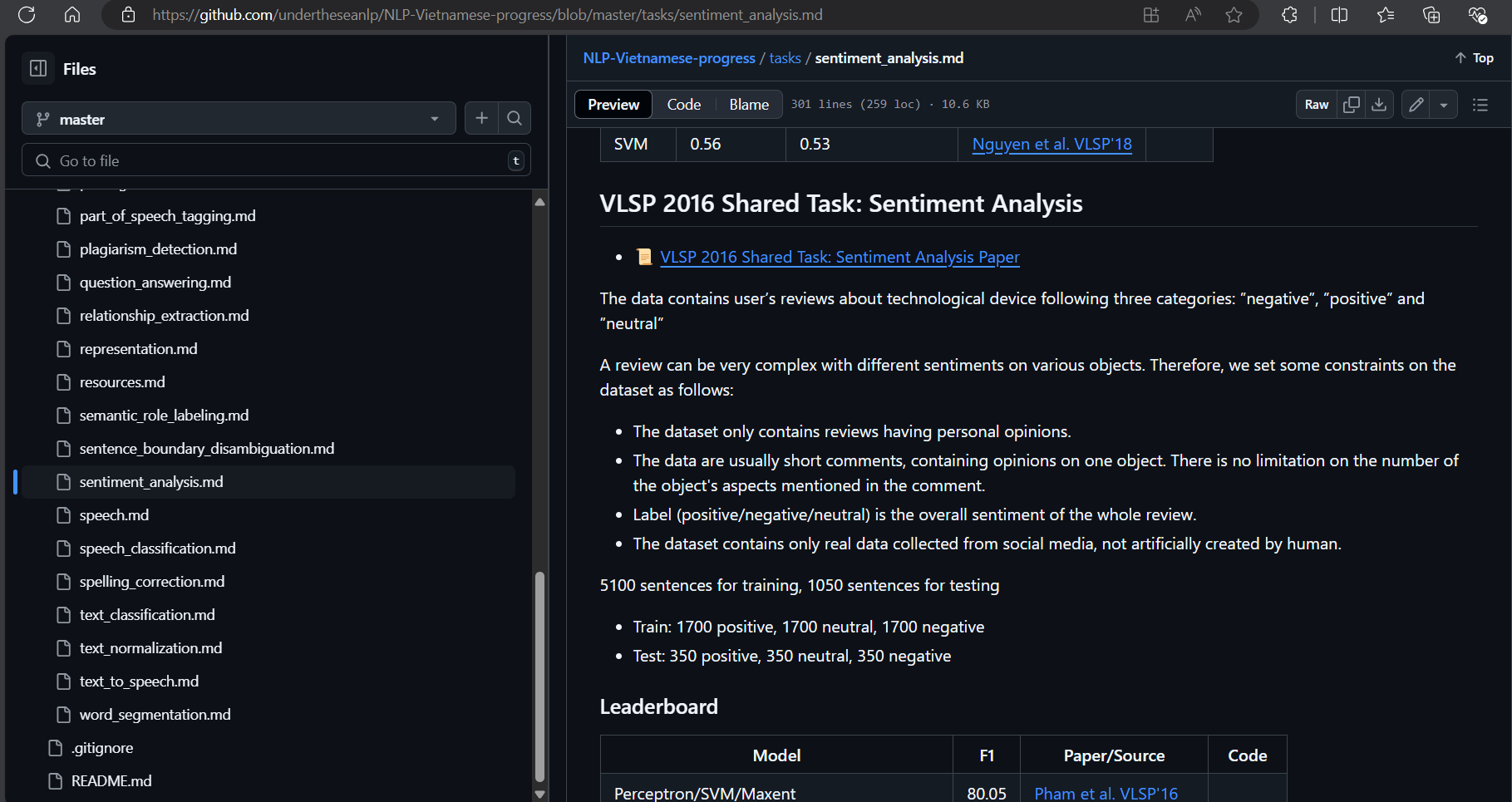
—----------------------

(P) Project: ML for NLP

Perform **sentiment analysis** using **SVM** and **Naive Bayes** and compare the results.

Dataset:

* <https://github.com/undertheseanlp/NLP-Vietnamese-progress/blob/master/tasks/sentiment_analysis.md?authuser=0>
* <https://www.kaggle.com/datasets/tuannguyenvananh/vietnamese-text-classification-dataset?authuser=0>



Paper: [2016 pham A Lightweight Ensemble Method for Sentiment.pdf - Google Drive](https://drive.google.com/file/d/1V87Z0Jjgt47nqcN8BaDLU5TLAqzUn1pN/view)

Note:

pip freeze > requirements.txt

pip install -r requirements.txt

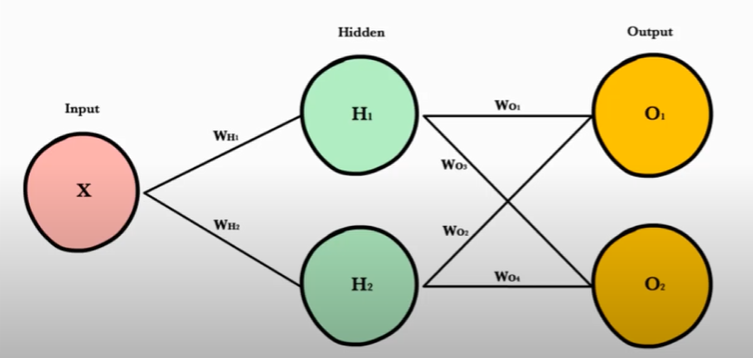
TensorFlow Neural Networks

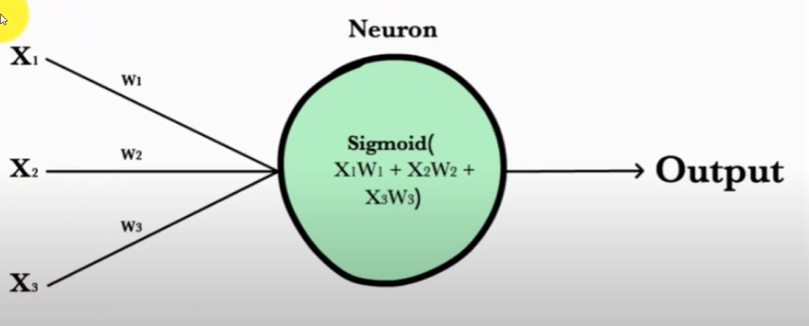
Material:

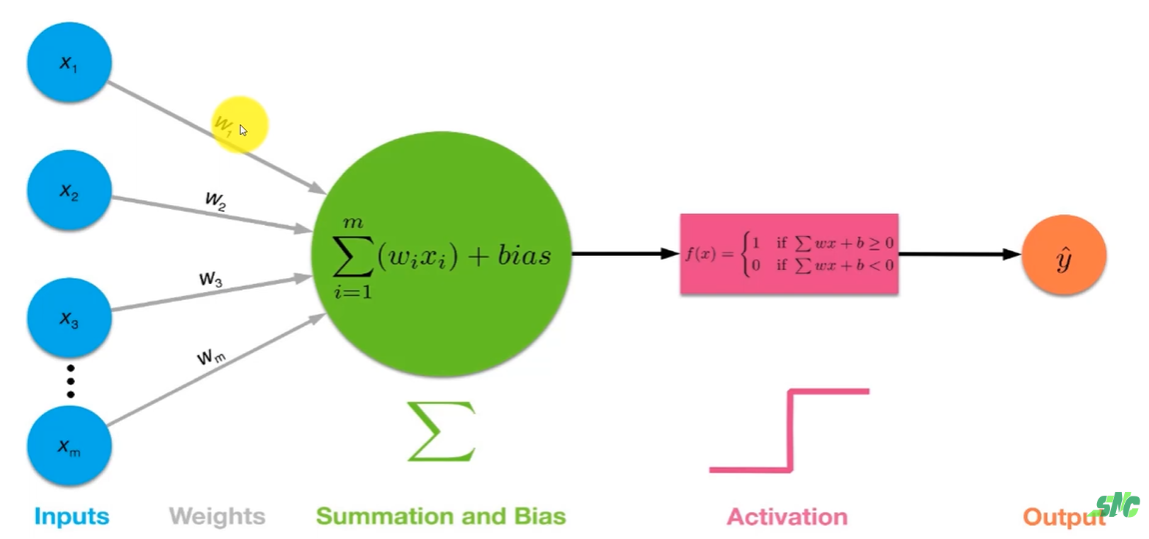
1.Deep Learning Book: [Deep Learning (deeplearningbook.org)](https://www.deeplearningbook.org/?authuser=0)

2.TensorFlow Tutorial: [Tutorials | TensorFlow Core](https://www.tensorflow.org/tutorials?authuser=0)

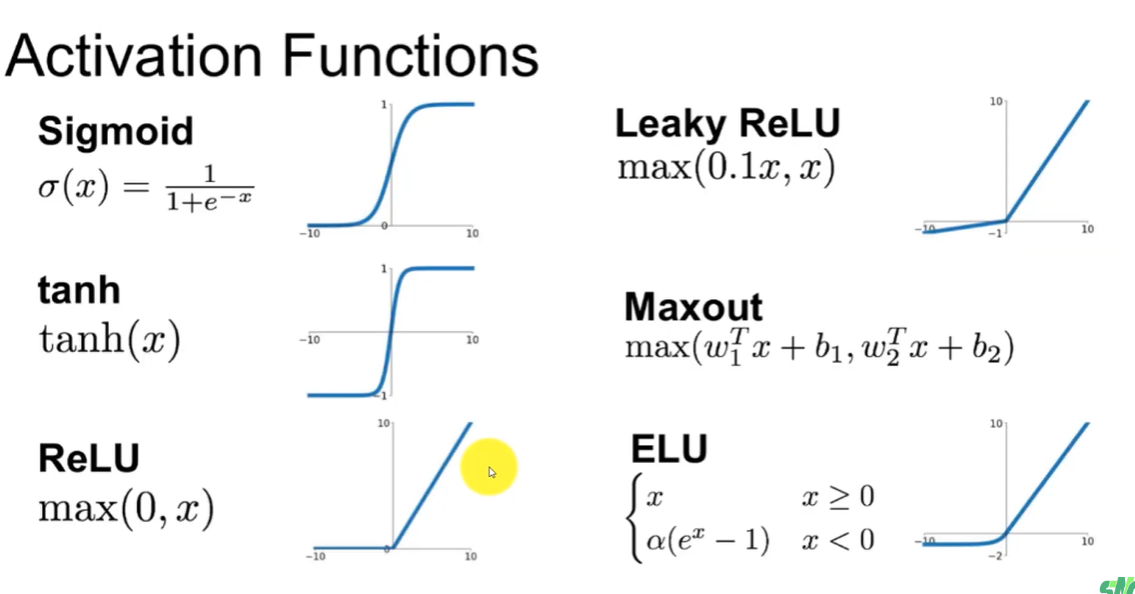
1. Neural Network







Activation Function (trả về giá trị 1 hoặc 0 => output)



* Hàm phi tuyến tính: Sigmoid, tanh, ReLU.

Loss Function

* Loss Function phổ biến: MSE (L2) and Cross-entropy Loss.

—----------------------------------

(P) Project: DL for NLP

Build a **simple neural network** for **text classification** and report the training process and results.

Dataset: <https://github.com/duyvuleo/VNTC?authuser=0>

Notes Week 4:

1. Các kỹ thuật **augmentation** trong NLP (xử lý khi số lượng mẫu của các lớp không cân bằng, vd topic A có nhiều mẫu x3 lần topic B => ….)

* Synonym Replacement (Thay thế từ đồng nghĩa)
* Thay thế một số từ trong câu bằng các từ đồng nghĩa
* Câu gốc: "Tôi rất vui khi được làm việc với bạn."
* => "Tôi rất **hạnh phúc** khi được làm việc với bạn."
* Random Insertion (Chèn từ ngẫu nhiên)
* "Tôi đi học mỗi ngày." => "Tôi **luôn** đi học mỗi ngày."
* Làm phong phú hơn, không thay đổi ngữ nghĩa chính.
* Random Deletion (Xóa từ ngẫu nhiên)
* Câu gốc: "Chúng tôi đã đến công viên vào buổi chiều."
* => "Chúng tôi đã đến công viên **chiều”**
* Back-Translation (Dịch ngược)
* Dịch câu sang một ngôn ngữ khác (tiếng anh, tiếng trung, tiếng nhật, tiếng đức,...) rồi dịch ngược lại tiếng Việt.
* Có thể dùng Google Translate
* Hướng dẫn tạo API\_KEY trên GG: [Generate Google Translate API Key - TranslatePress](https://translatepress.com/docs/automatic-translation/generate-google-api-key/#:~:text=Setting%20up%20API%20keys%20for%20Google%20Translate%201,add%20Google%20Translate%20API%20Key%20in%20TranslatePress%20)
* "Tôi thích đọc sách vào buổi tối." => "I like reading books in the evening." => "Tôi thích đọc sách vào **buổi đêm**."
* Augmentation Sử Dụng PhoBERT (BERT-based Augmentation)
* "Tôi đang học lập trình." => Sử dụng PhoBERT để dự đoán từ thay thế: "Tôi đang **nghiên cứu** lập trình."
* Tương tự mặt ngữ nghĩa, khác cách diễn đạt
* Word Swap:
* "Cô ấy chạy rất nhanh." => "Cô ấy **rất nhanh** chạy."