



# CORPORATE STOCK REVIEW !



Introduction to Artificial Intelligence - IT3160E



Nguyen Chi Long - Ngo Xuan Bach  
Le Xuan Hieu - Dinh Viet Quang

# PROBLEM STATEMENT

Stock analysis, which has always been a crucial aspect of any organization, is a method for investors and traders to make buying and selling decisions.

By studying and evaluating past and current data, investors and traders attempt to gain an edge in the markets by making informed decisions

## OUR TARGETS

Evaluate how the sentiment impacts stock performance through article text.

Assign a large set of corporate stock reviews to the several levels of polarity of opinion as accurately as possible.

[Back to Agenda Page](#)



# APPROACH

Sentiment Analysis Techniques

1	Explore the collected data
2	<b>Input:</b> sentences of articles or contents from an URL link <b>Output:</b> sentiment classification.
3	Evaluation metrics: accuracy and confusion matrix
4	Compare the results of each model and find the best model
5	Load the models for demonstration and deployment

# Data Gathering Phases

Scrapy



1

Data crawling

Doccano



2

Data annotation

3

Data processing and  
Model training



Python

Demo and deployment

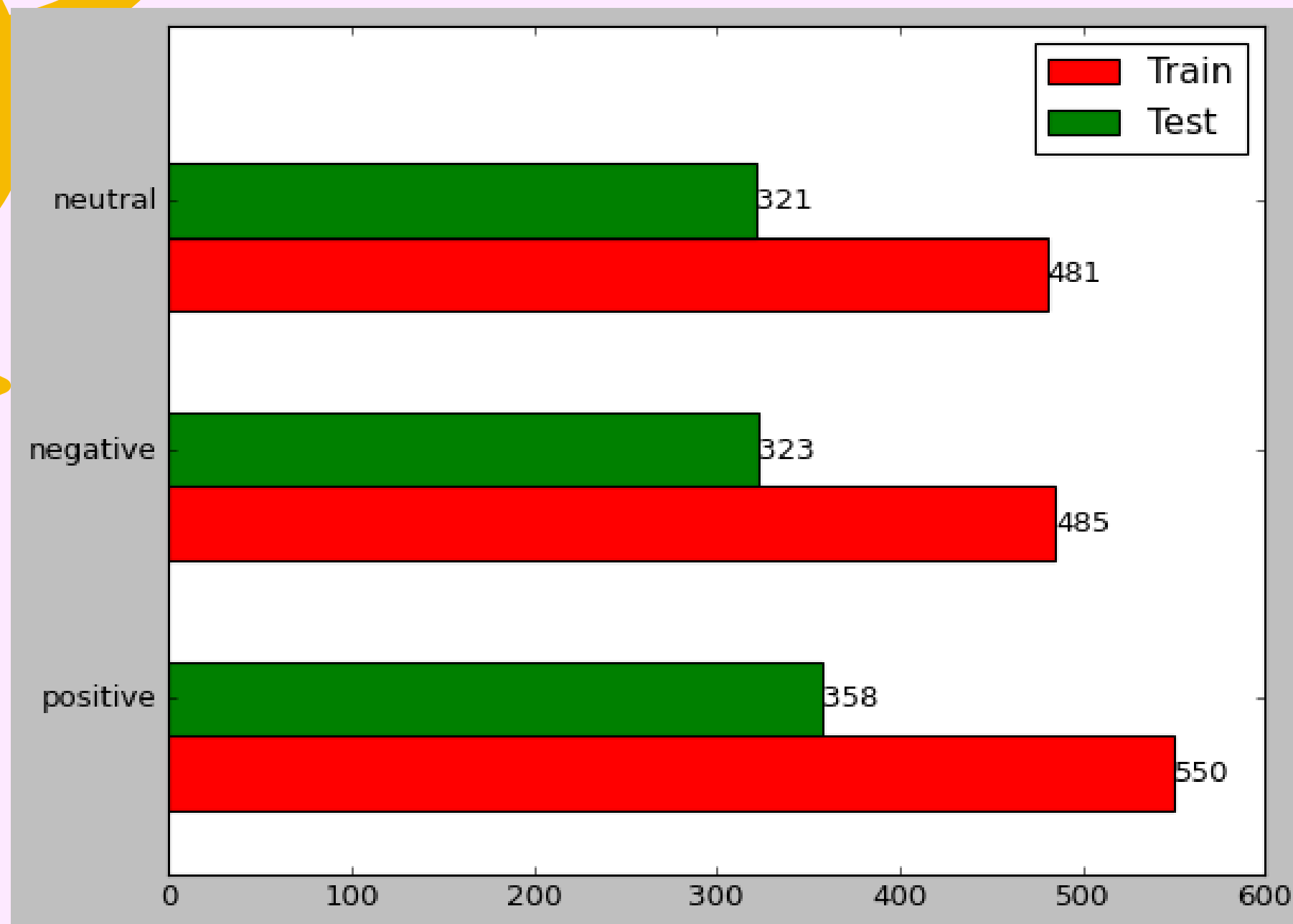
jupyter



Flask

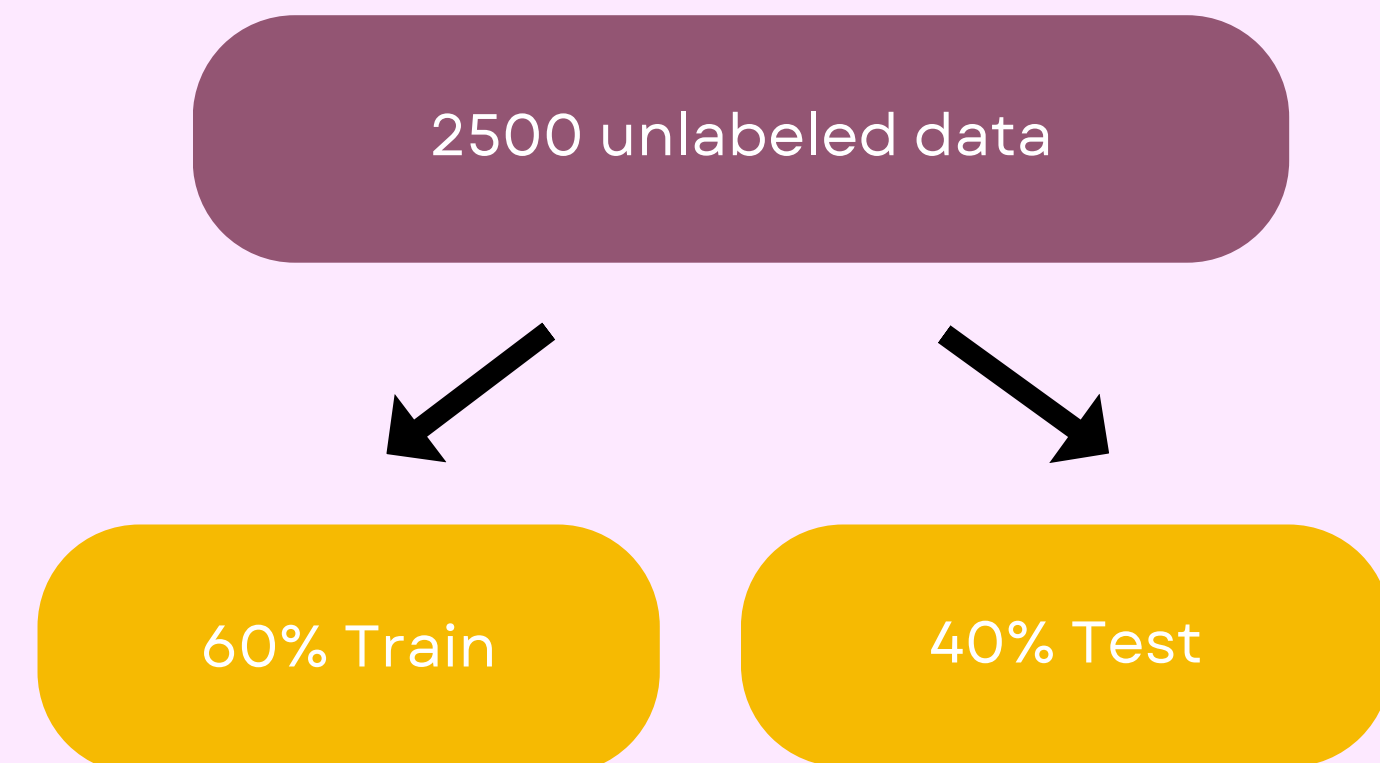
4

[Back to Agenda Page](#)



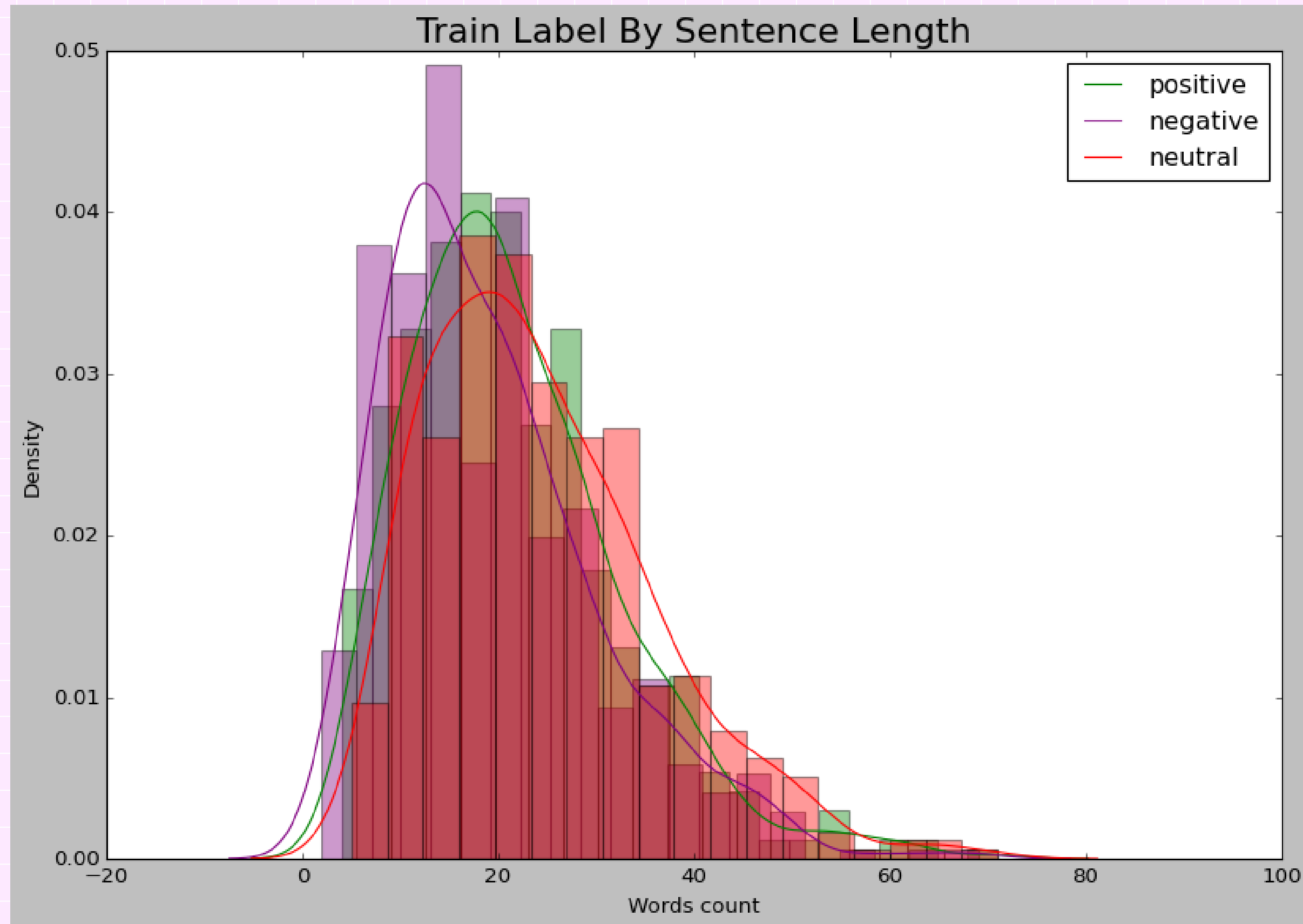
# OUR DATASET

Collect corporate stock reviews from various sources (CafeF, VnExpress etc.)



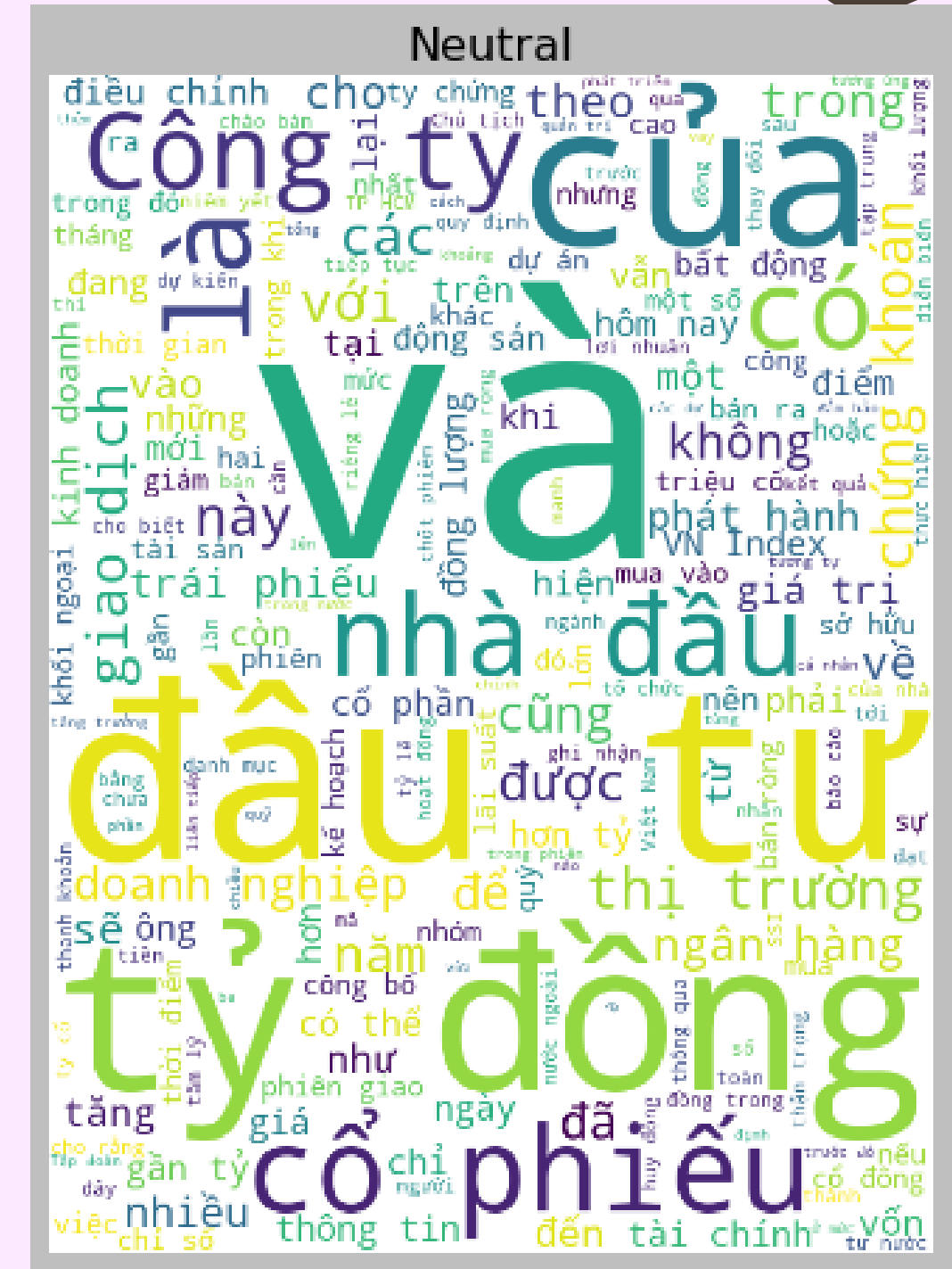
# SENTENCE LENGTH

Good  
Job!!





# WORD FREQUENCY



[Back to Agenda Page](#)



# DATA PREPROCESSING

## TF - IDF



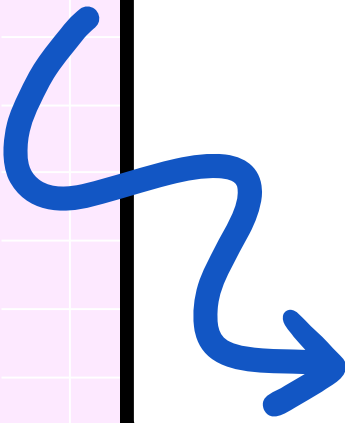
$$\text{TFIDF}(t, d, N) = \text{tf}(t, d) \times \log(N/\text{df}(d, t)+1)$$



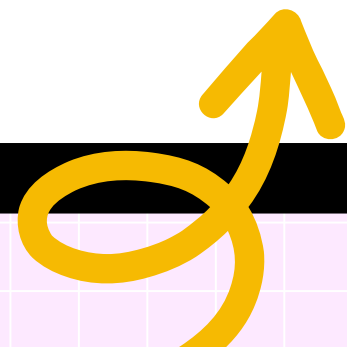
TF - IDF

TERM  $t$  WITHIN DOCUMENT  $d$

$\text{tf}(t, d)$  = frequency of  $t$  in  $d$   
 $\text{df}(d, t)$  = number of documents  $d$  containing  $t$ .  
 $N$  = total number of documents.



Term frequency - Inverse document frequency: is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.





# MODELING

## Random Forest

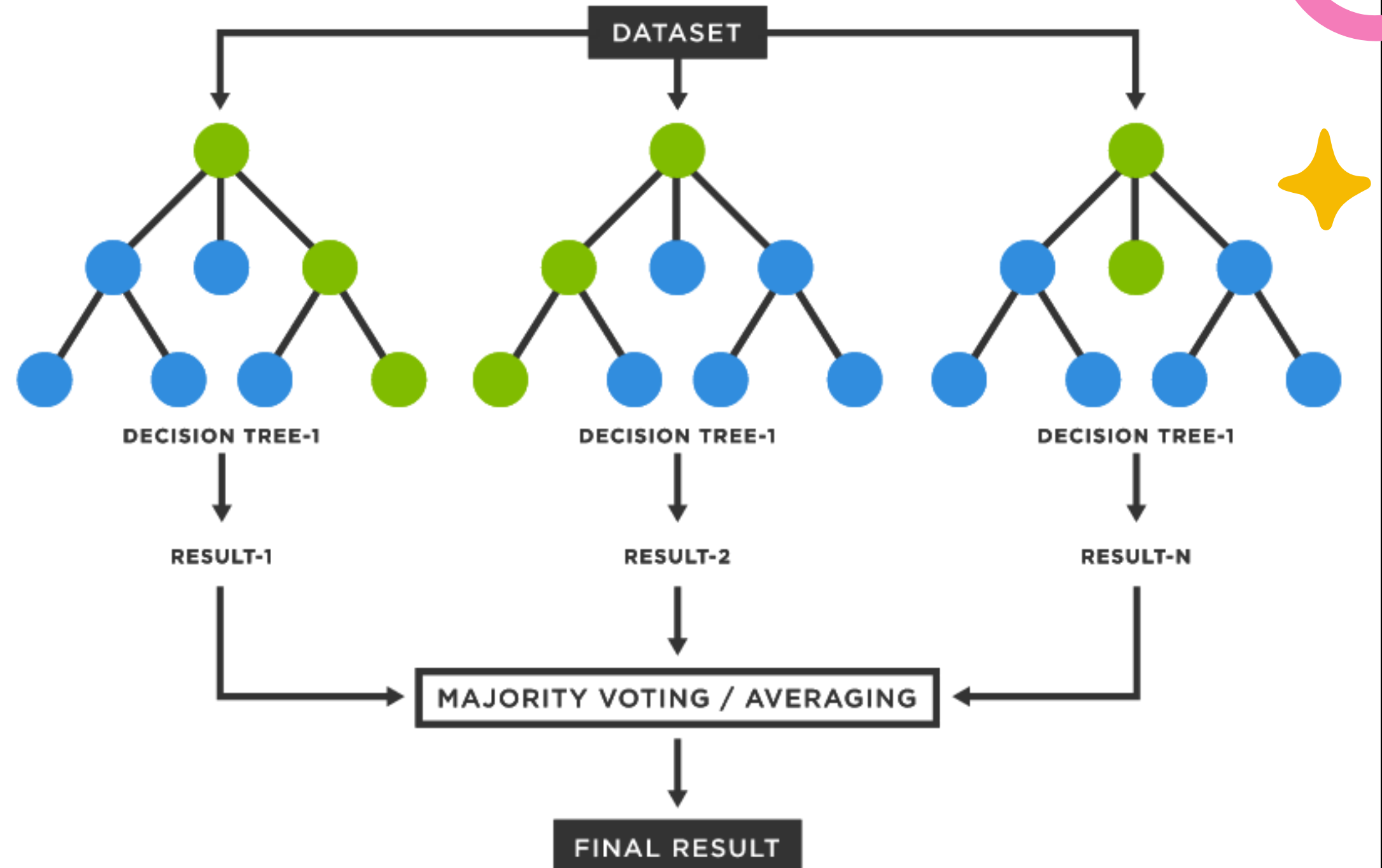
Grows multiple classifier trees with bootstrap sampling with replacement to train each tree with a different part of the dataset.

Three main hyperparameters:

- + Depth of the trees: 50.
- + Max features: "sqrt".
- + Random state

Criterion = 'Gini'

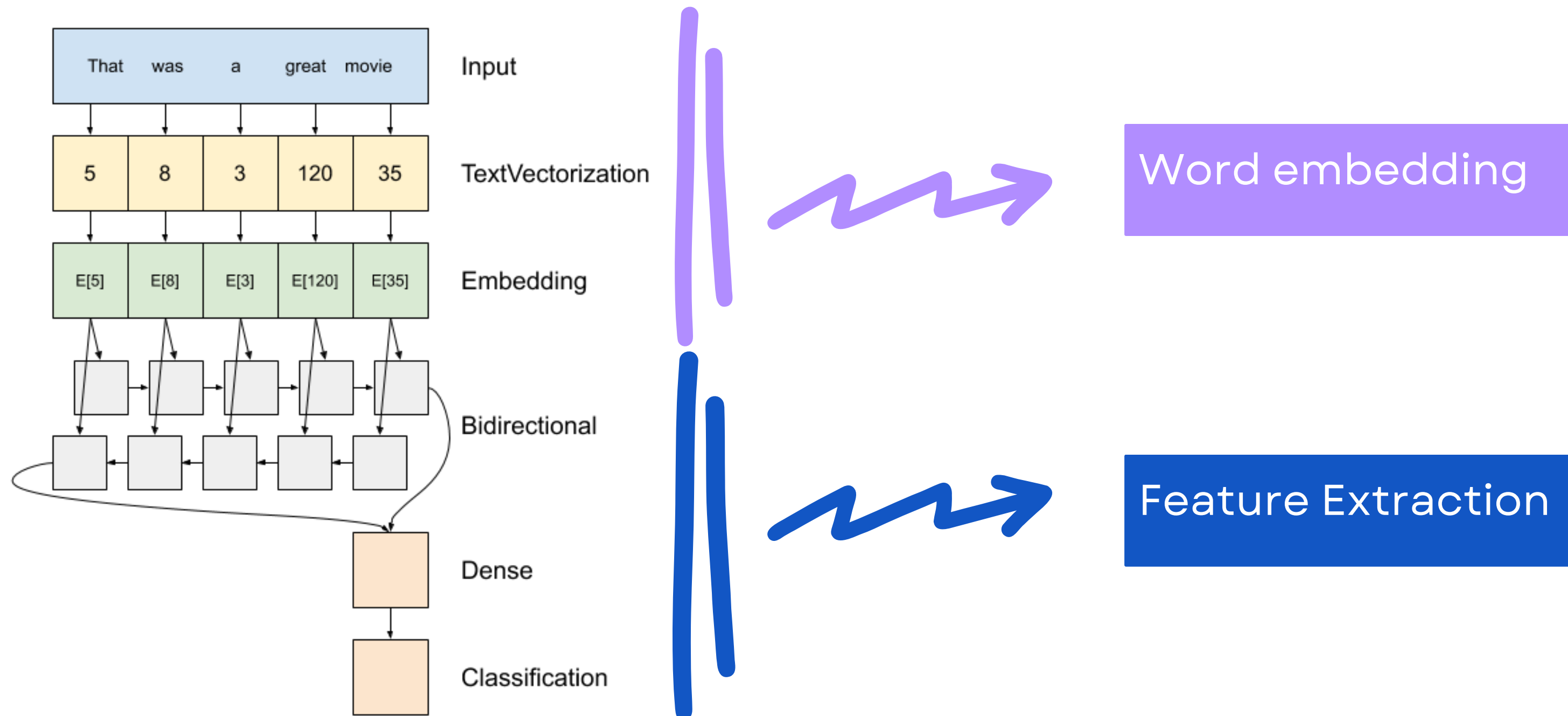
[Back to Agenda Page](#)



# MODELING

## Neural Network

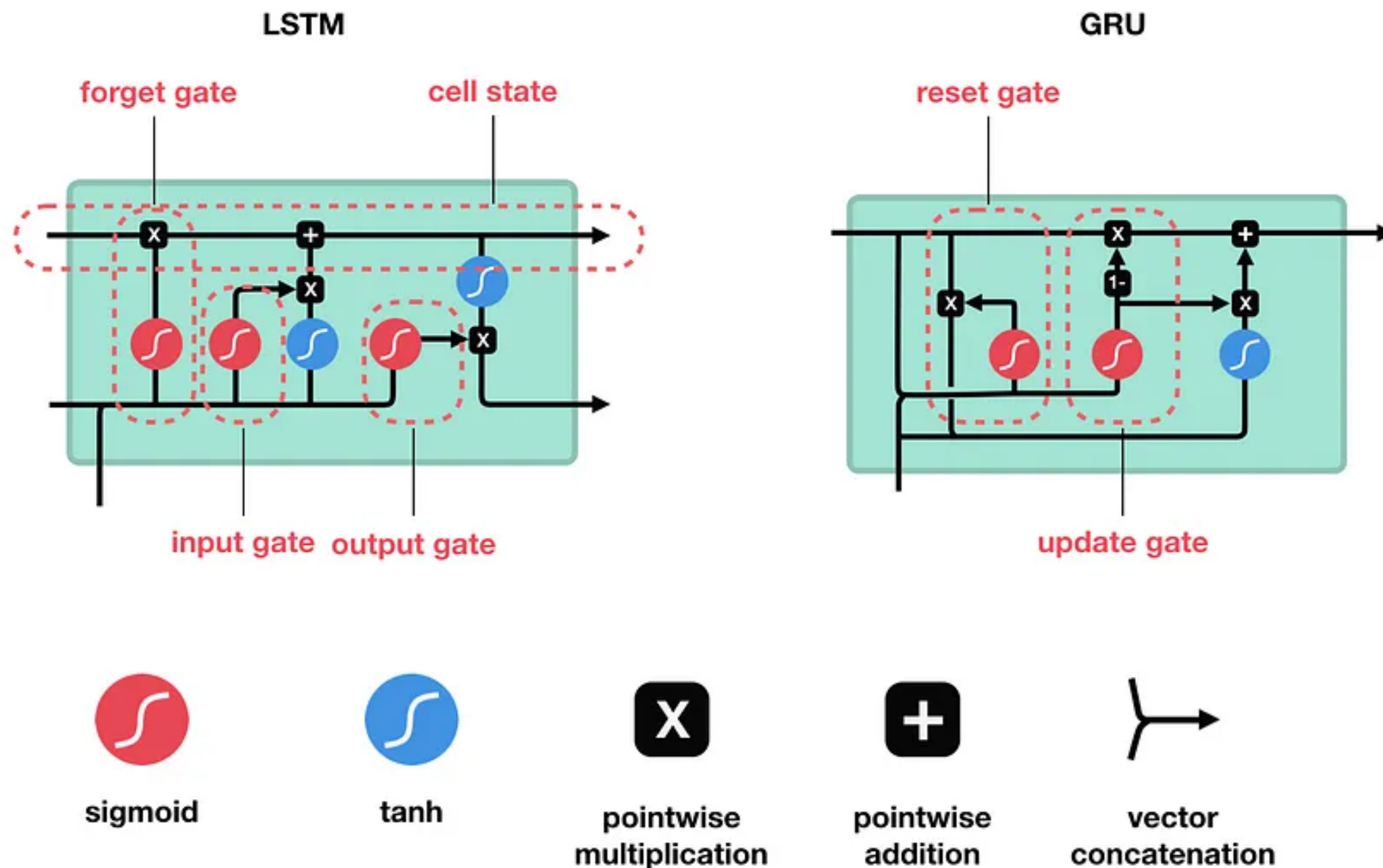
### Deep Bidirectional Architecture



# MODELING

## Neural Network

### Long Short-term Memory (LSTM) and Gated Recurrent Unit (GRU)



Newer generation of Recurrent Neural networks.

The core concept of LSTM are the cell state, and its various gates.

LSTM cell replacement: GRU (Similar to an LSTM cell).

GRU's got rid of the cell state and used the hidden state to transfer information.

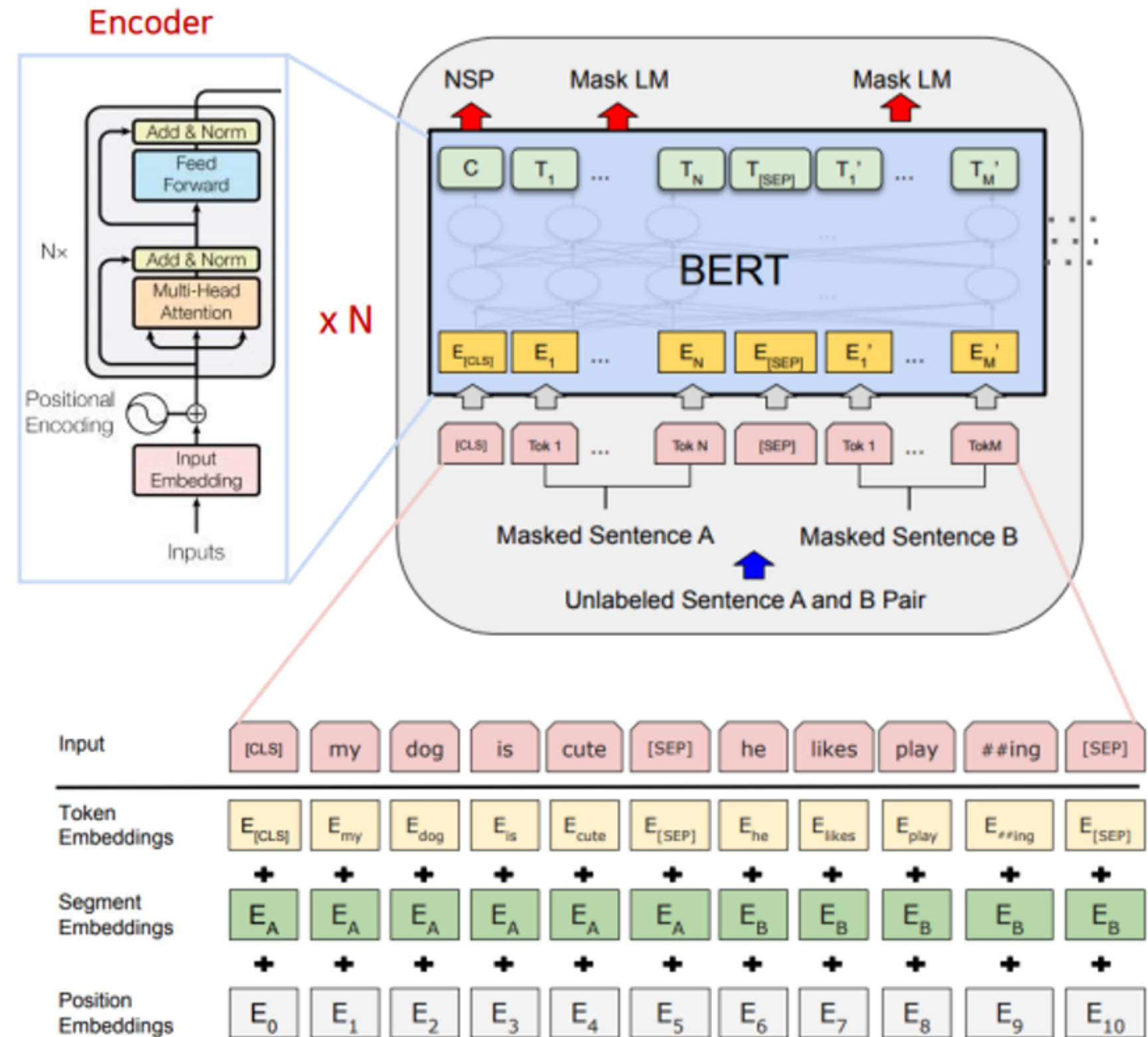
# MODELING

## phoBERT Architecture

- 1 Embed the dataset using **3 kinds of embedding**.
- 2 Fine-tuning PhoBERT on a large Vietnamese corpus.
- 3 Use the model for evaluation.



Back to Agenda Page



# MODELING

## XLM - Roberta

1

3 pre-trained language modelling objectives are used: CLM, MLM, TLM.

2

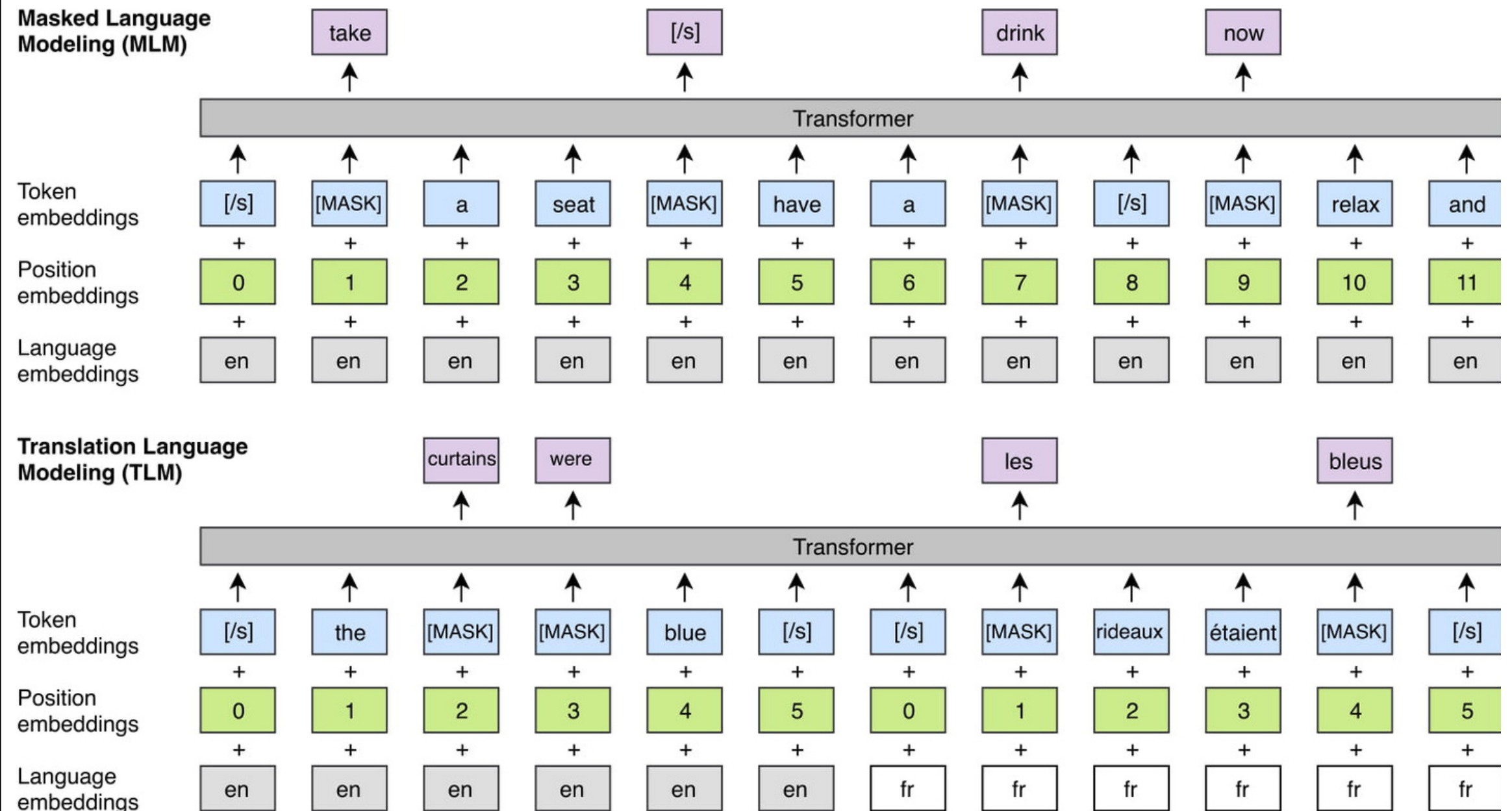
The model is pre-trained over RoBERTa architecture.

3

Use the model for evaluation.



[Back to Agenda Page](#)





# EVALUATIONS: QUANTITATIVE RESULTS

Model	Train Accuracy (%)	Test Accuracy (%)	Inference Time	Input Fields
PhoBERT	99.86	81.24	7.43h	Text Fields Only
RF + TF - IDF	99.61	68.76	3.2s	Text Fields Only
DBLSTM	99.60	62.28	300.1s	Text Fields Only
DBGRU	99.80	63.67	412.4s	Text Fields Only
XLM - RoBERTa	<b>99.92</b>	<b>82.04</b>	3.30h	Text Fields Only

The classification results for each model

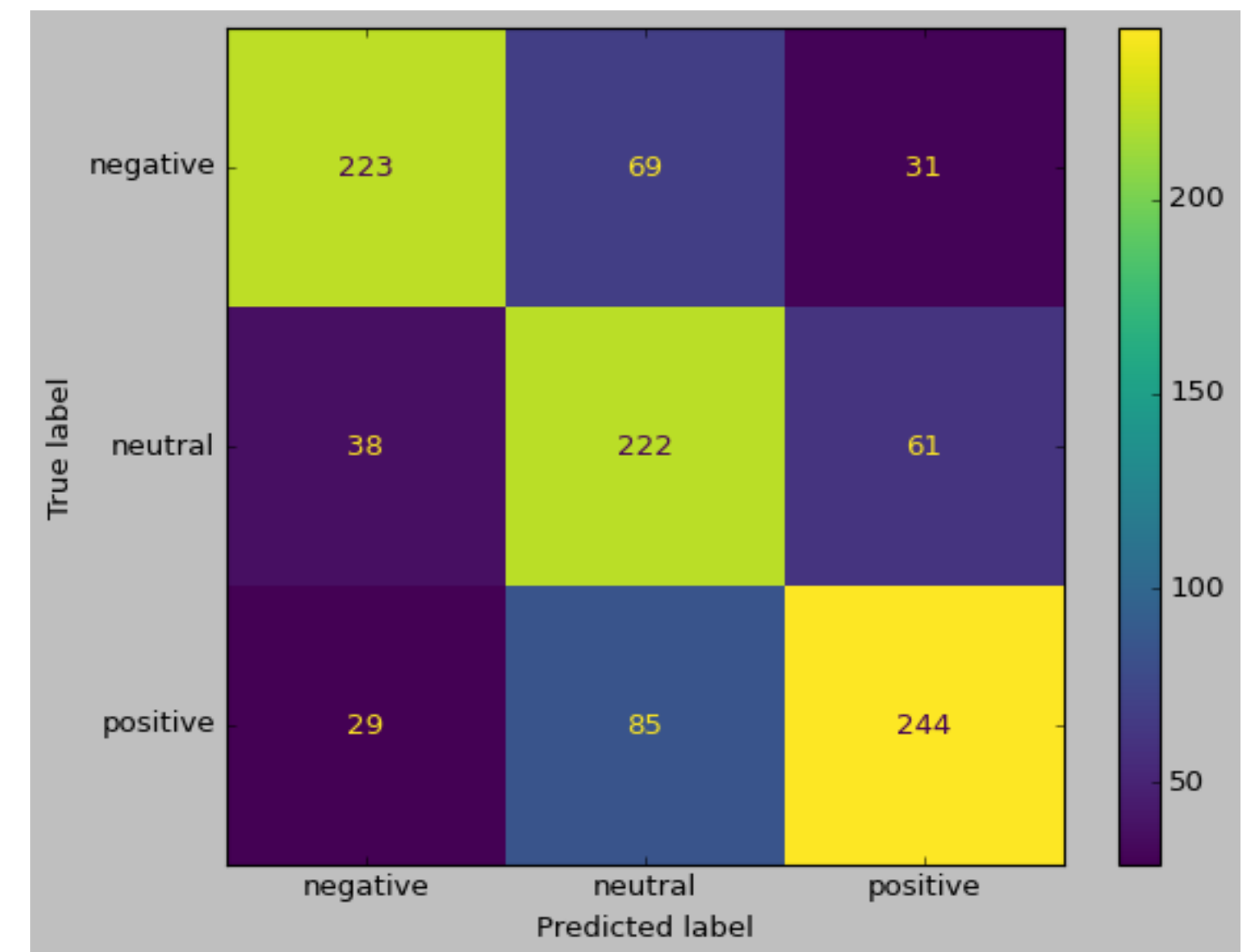
[Back to Agenda Page](#)

# EVALUATIONS: QUALITATIVE RESULTS

What Random Forest model have learned and classified through our datasets.

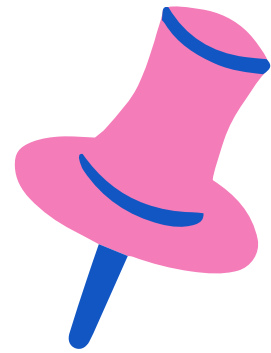
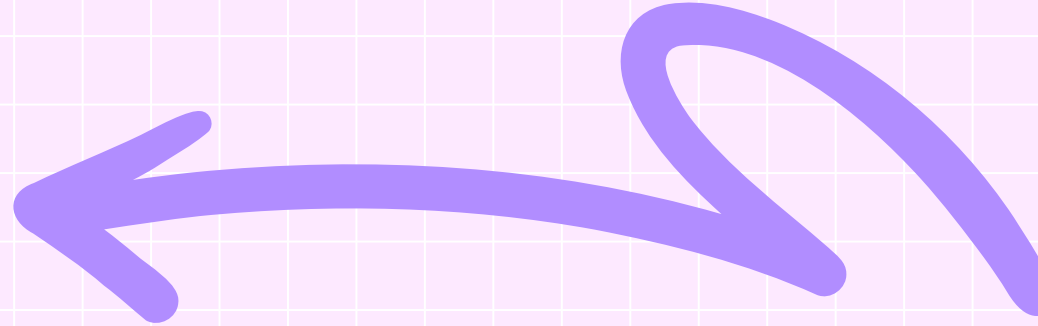
	precision	recall	f1-score	support
negative	0.77	0.69	0.73	323
neutral	0.59	0.69	0.64	321
positive	0.73	0.68	0.70	358
accuracy			0.69	1002
macro avg	0.70	0.69	0.69	1002
weighted avg	0.70	0.69	0.69	1002

Train Accuracy: 0.996042216358839  
Test Accuracy: 0.687624750499002






# DIFFICULTIES AND LIMITATIONS



## Difficulties

Do not have common knowledge about the problem.

Do not find the effective way to preprocess the data.



## Limitations

Limited hardware resources.

Limited knowledge about the theoretical properties of the models



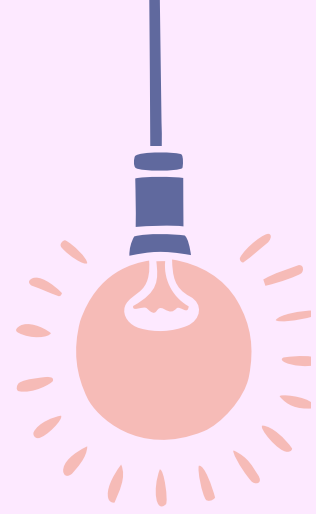
# CONCLUSIONS

## SUMMARY

- By using various methods and tools, we have trained five different models on our dataset.
- **The best model among all is XLM - RoBERTa.**
- To gain success in solving a problem, having great insight about the problem is crucial.

## POSSIBLE EXTENSIONS

- Understanding the dataset and collect more data for training accuracy, finding better training strategy for each models.
- Research and implement new models like ViDeBERTa.
- Try out the application of self-supervised and semi-supervised learning methods for this problem.



6 THANKS FOR  
WATCHING

