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.
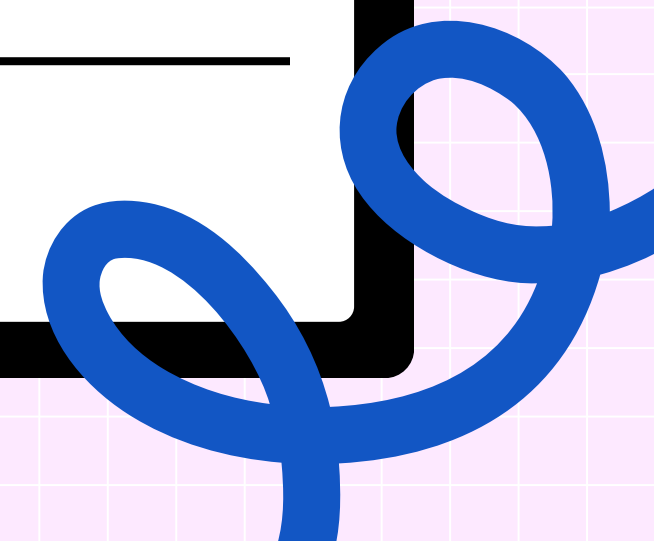
Back to Agenda Page

APPROACH

Sentiment Analysis Techniques

1	Explore the our collected data
2	Evaluation metrics: accuracy and confusion maxtrix
3	Compare the results of each model and find the best model
4	Load the models for demonstration and deploment

Data Gathering Phases

Scrapy

Demo and deployment

Doccano

3 Data processing and Model training

1 Data crawling

2 Data annotation

4
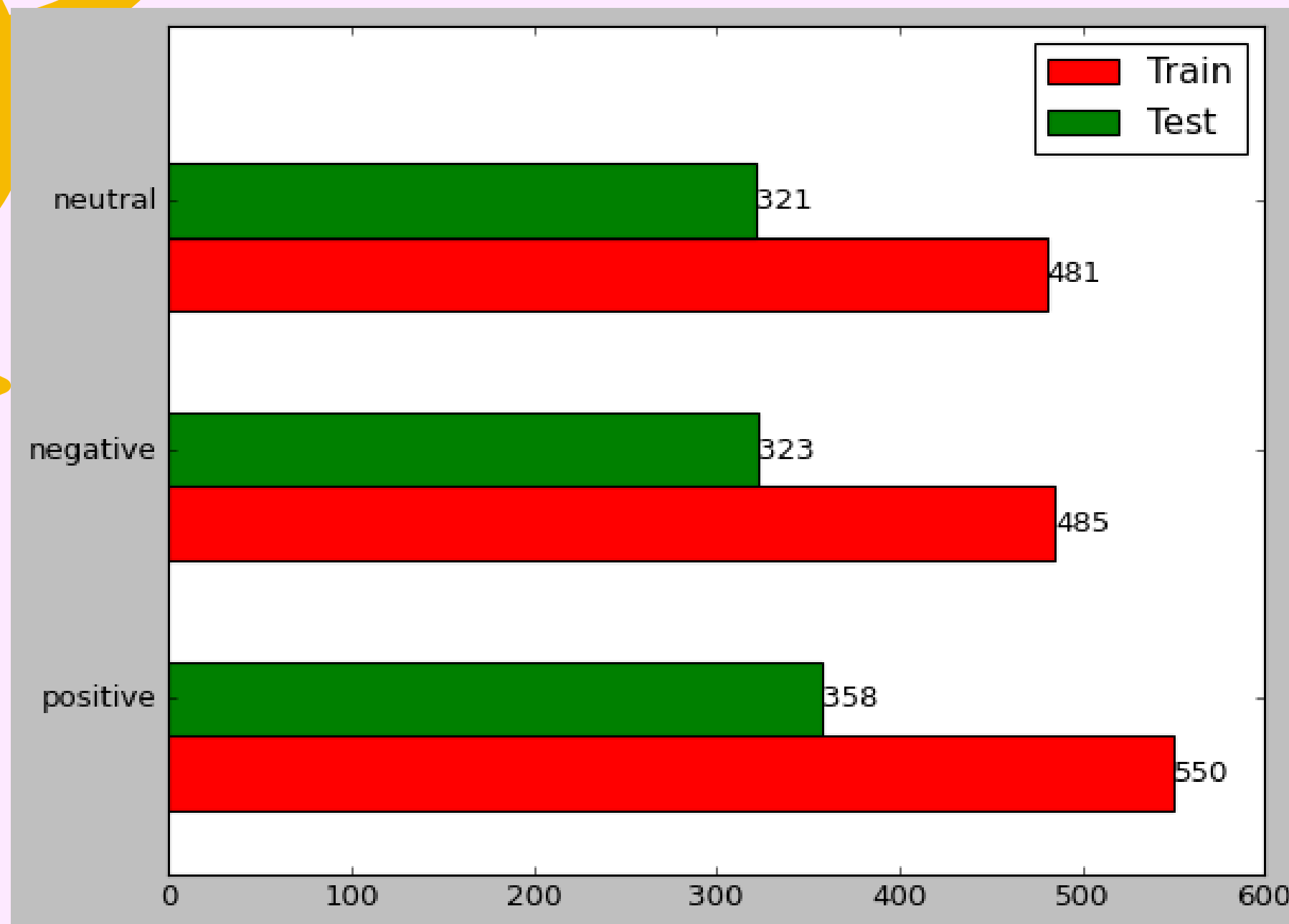
Python

OUR DATASET

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

The dataset contains 2500 additional unlabeled data. The number of positive, negative and neutral sentiment is more 800

Data split: 60% Train - 40% Test

Back to Agenda Page

EXAMPLE

"Positive" review: "Nhiều mã ngành này lọt vào nhóm góp mức tăng lớn nhất cho toàn thị trường, thị giá kịch trần ngay từ đầu buổi sáng"

"Negative" review: "Trong thông báo hôm nay, lãnh đạo FLC đã trình bày một loạt khó khăn khiến doanh nghiệp không thể công bố các thông tin theo đúng quy định sau khi cựu Chủ tịch HĐQT Trịnh Văn Quyết bị bắt cuối tháng 3/2022"

"Neutral" review: "Đóng cửa phiên giao dịch ngày 20/4, cổ phiếu NT2 có giá 29.400 đồng/cổ phiếu"
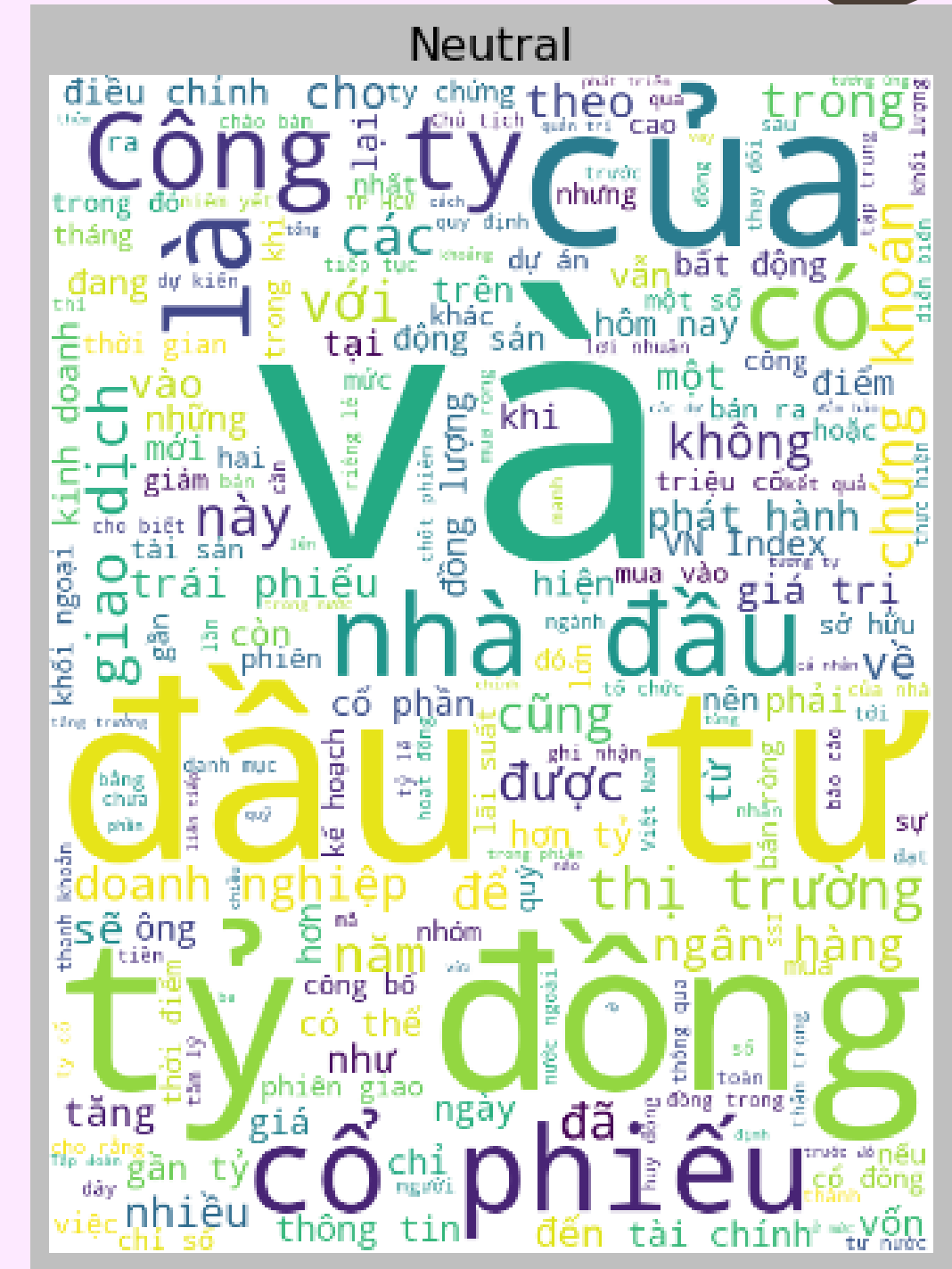
Back to Agenda Page

Train Label By Sentence Length

WORD FREQUENCY

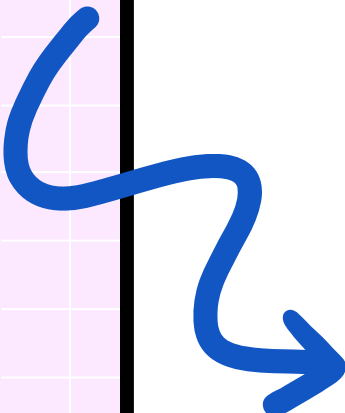Positive / Negative / Neutral word clouds
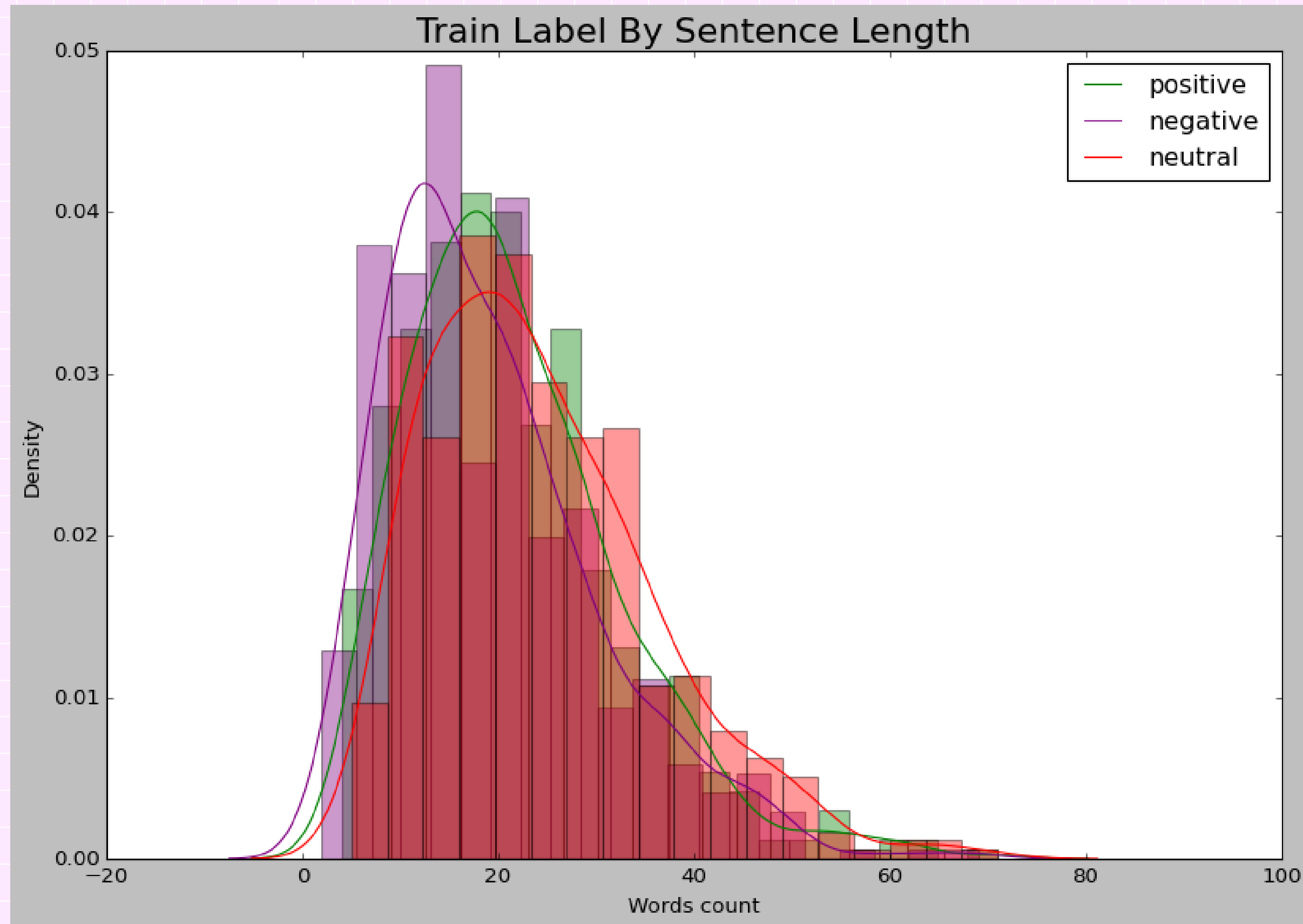
Back to Agenda Page

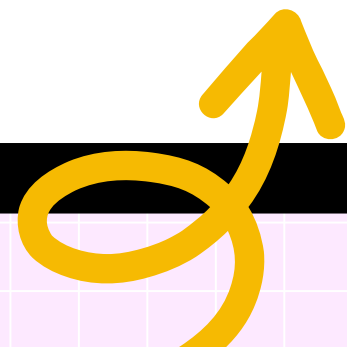# DATA PREPROCESSING

TF - IDF

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

TF - IDF
TERM t WITHIN DOCUMENT d

tf(t,d) = frequency of t in d
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.
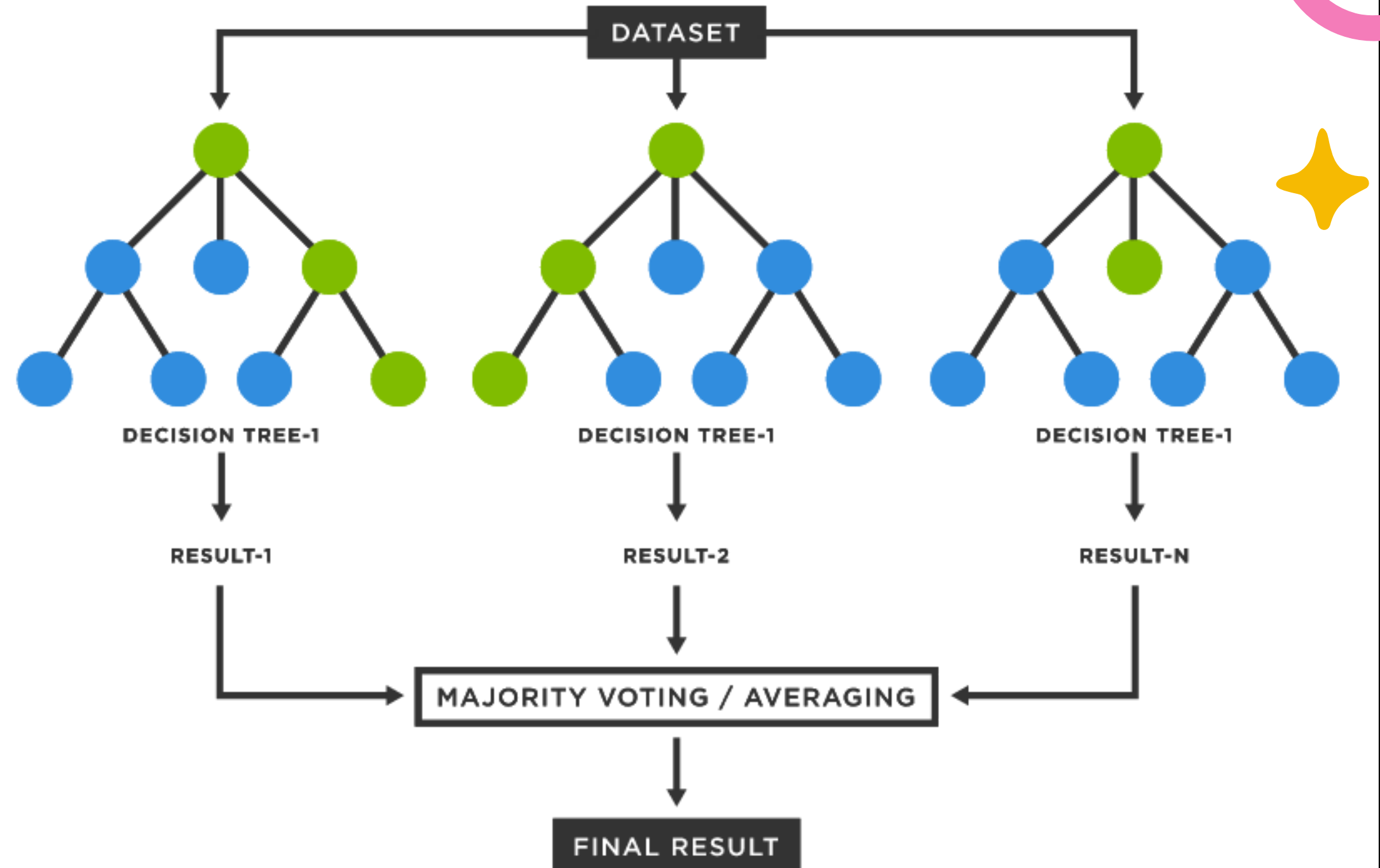
Random Forest

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

PREPROCESSING: WORD EMBEDDING

| That | was | a | great | movie | **Input** |

Lower and strip punctuation, turn word into index

| 5 | 8 | 3 | 120 | 35 | **TextVectorization** |

Turn positive integers (indexes) into dense vectors of fixed size
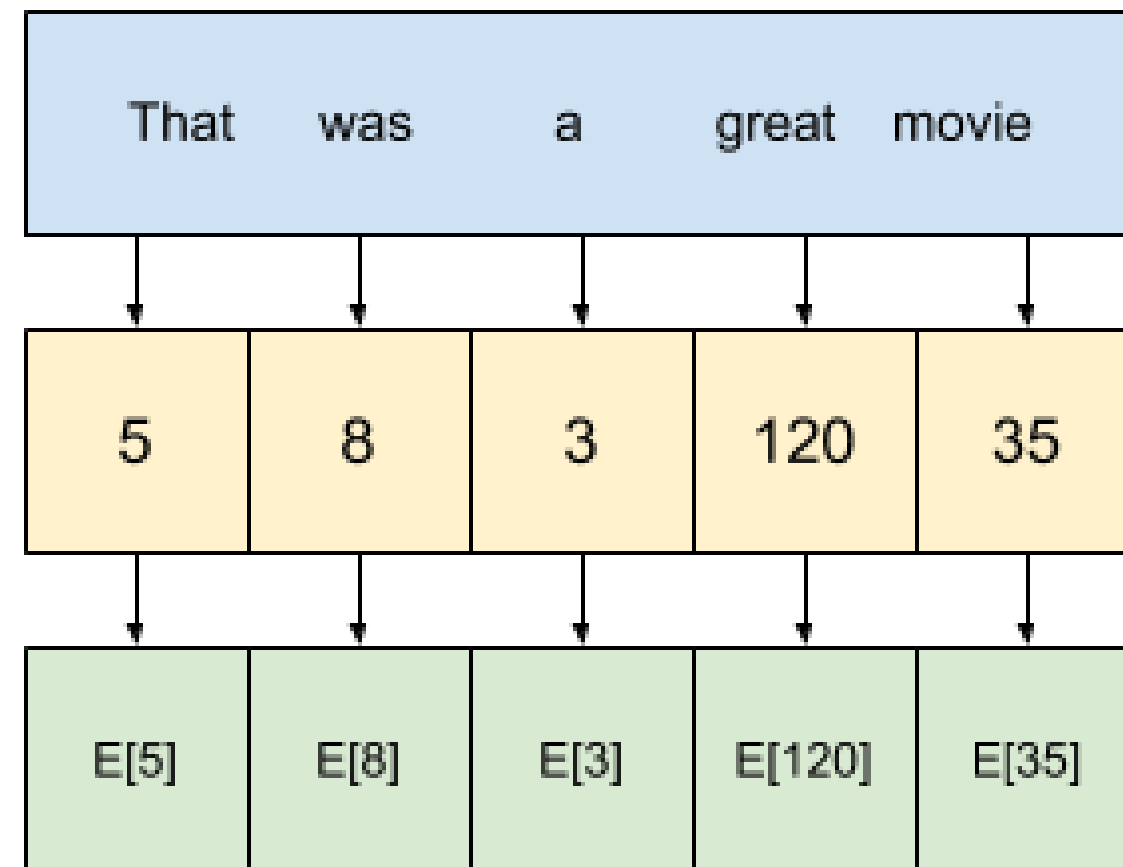
| E[5] | E[8] | E[3] | E[120] | E[35] | **Embedding** |

Word embeddings are a form of word representation designed to capture semantic relationships, wherein words that are closer together in the vector space are generally assumed to have similar meanings

MODELING Neural Network

Deep Bidirectional Architecture

| That | was | a | great | movie | Input |

| 5 | 8 | 3 | 120 | 35 | TextVectorization |

| E[5] | E[8] | E[3] | E[120] | E[35] | Embedding |

Bidirectional

Dense

Classification

Word embedding

Feature Extraction
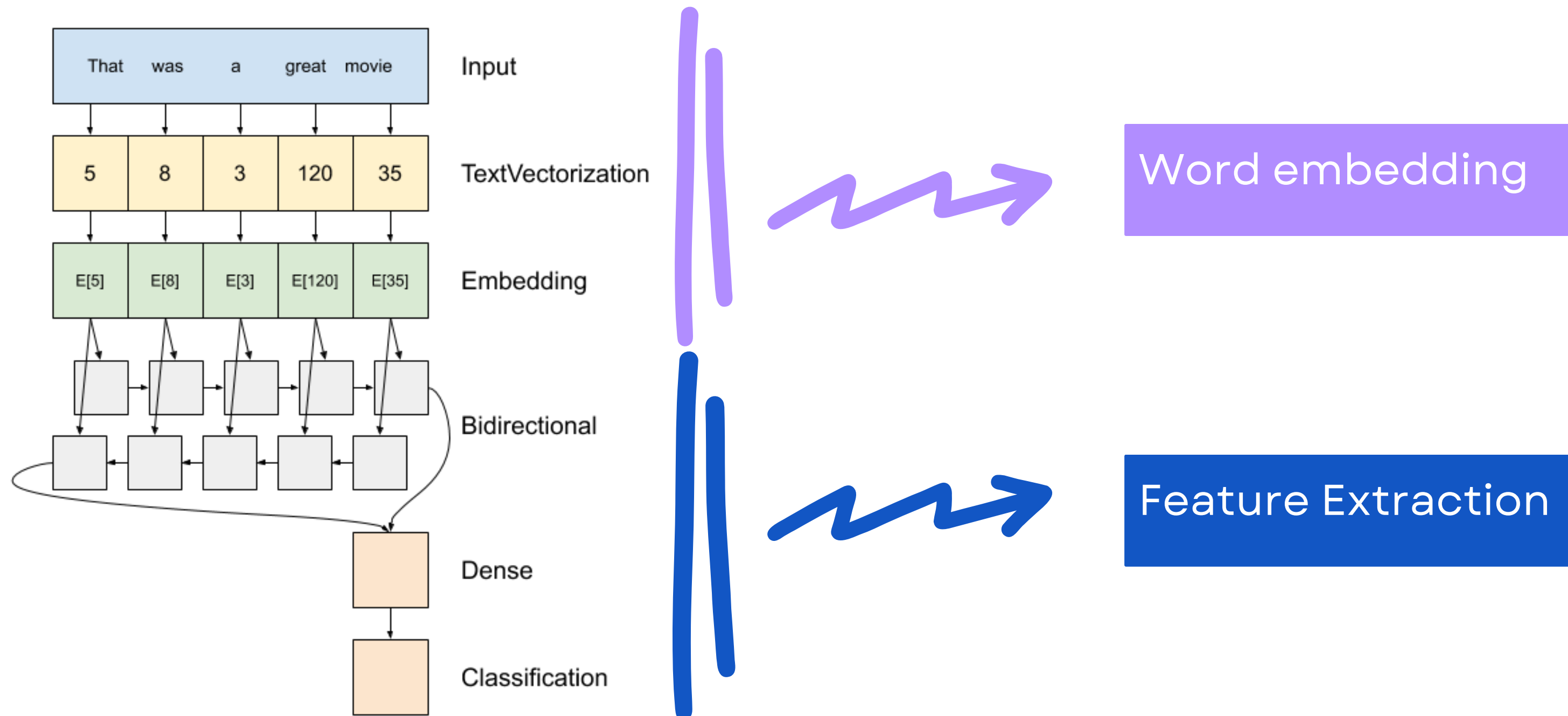
MODELING Neural Network

Long Short-term Memory (LSTM) and Gated Recurrent Unit (GRU)
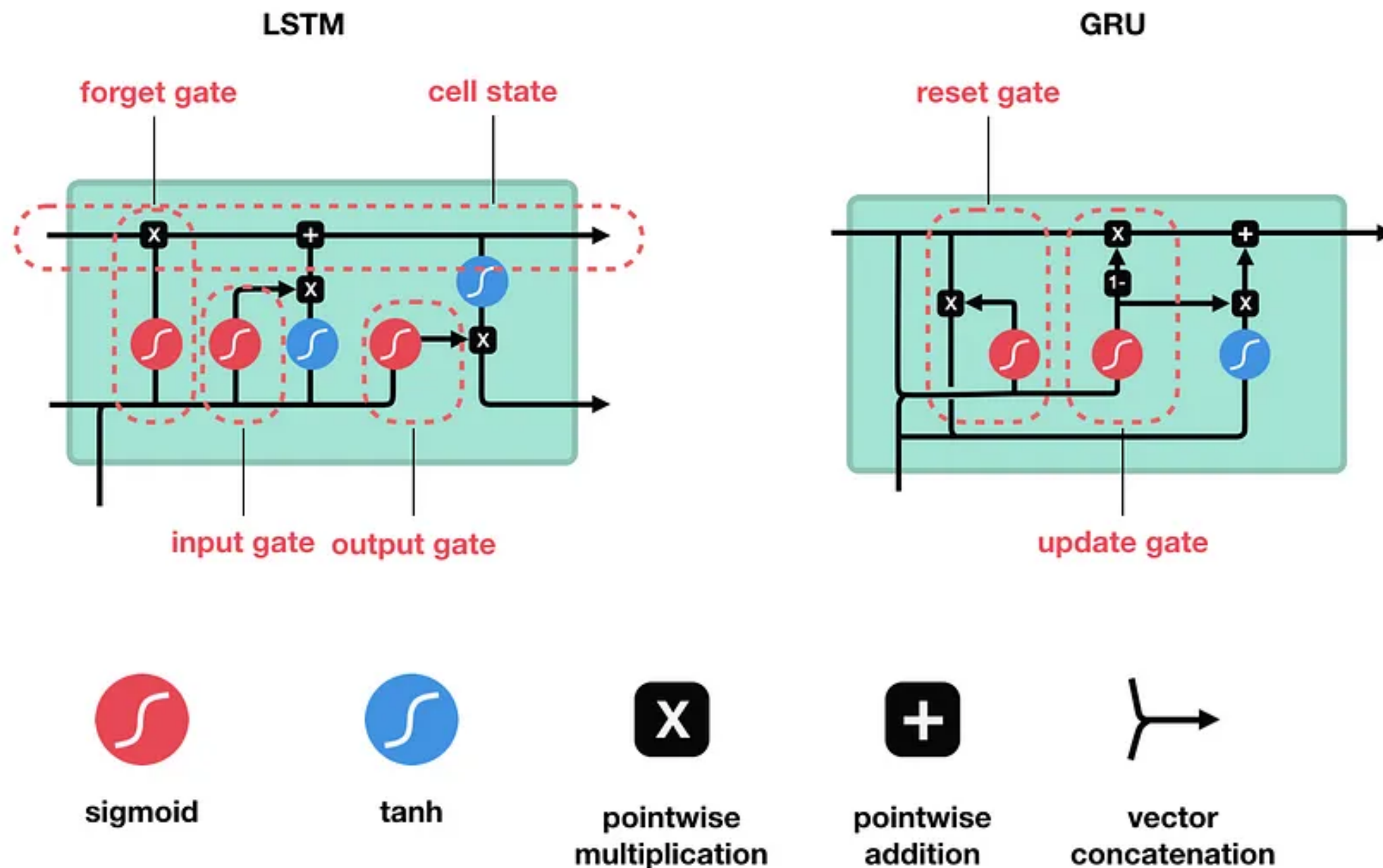
Newer generation of Recurrent Neural networks.

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 Archiecture

1 Build the dataset containing **3 kinds of sentiment.**

2 Fine-tuning PhoBERT with the dataset above.

3 Use the model for evaluation.

Back to Agenda Page

MODELING

XLM - Roberta

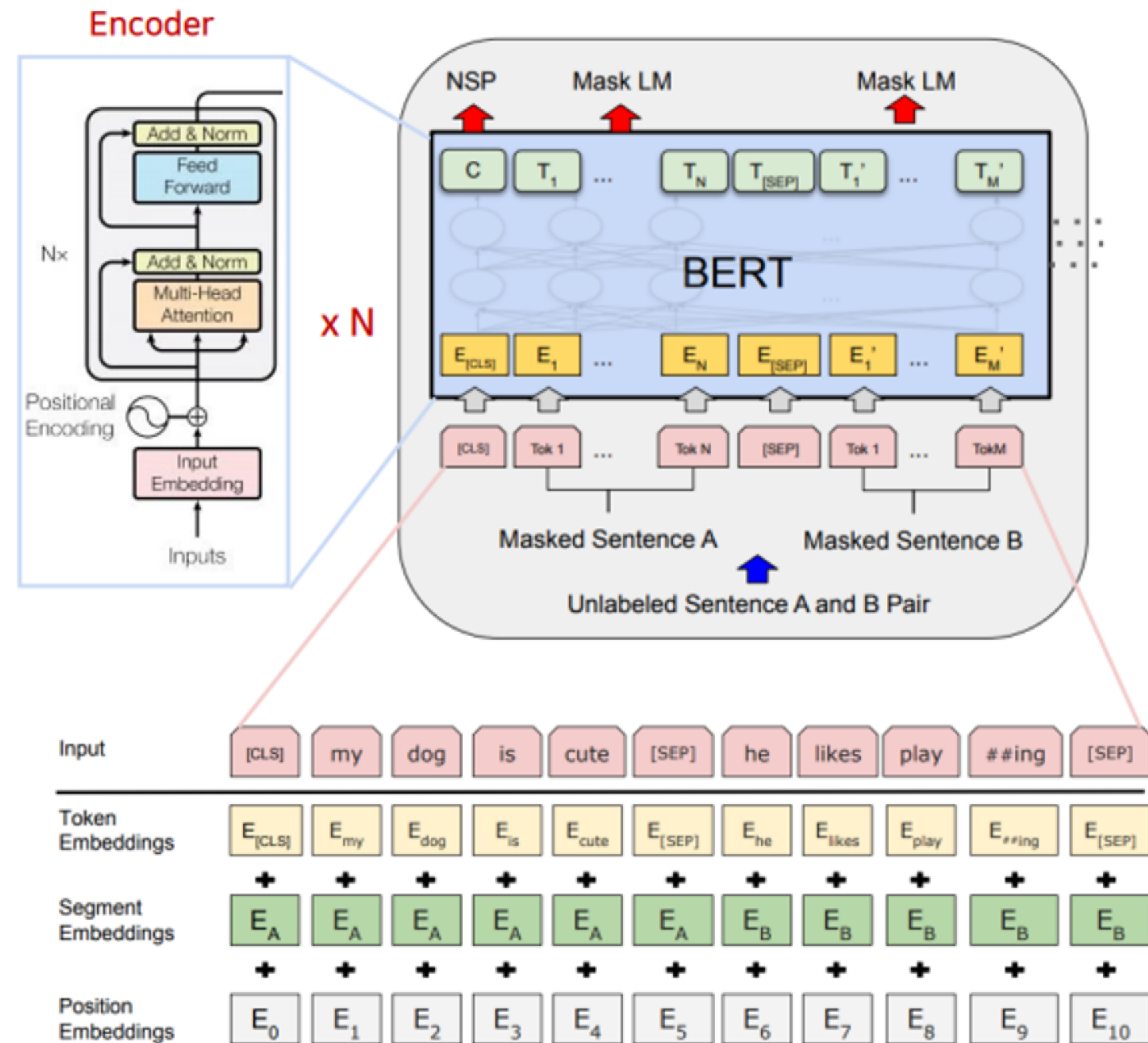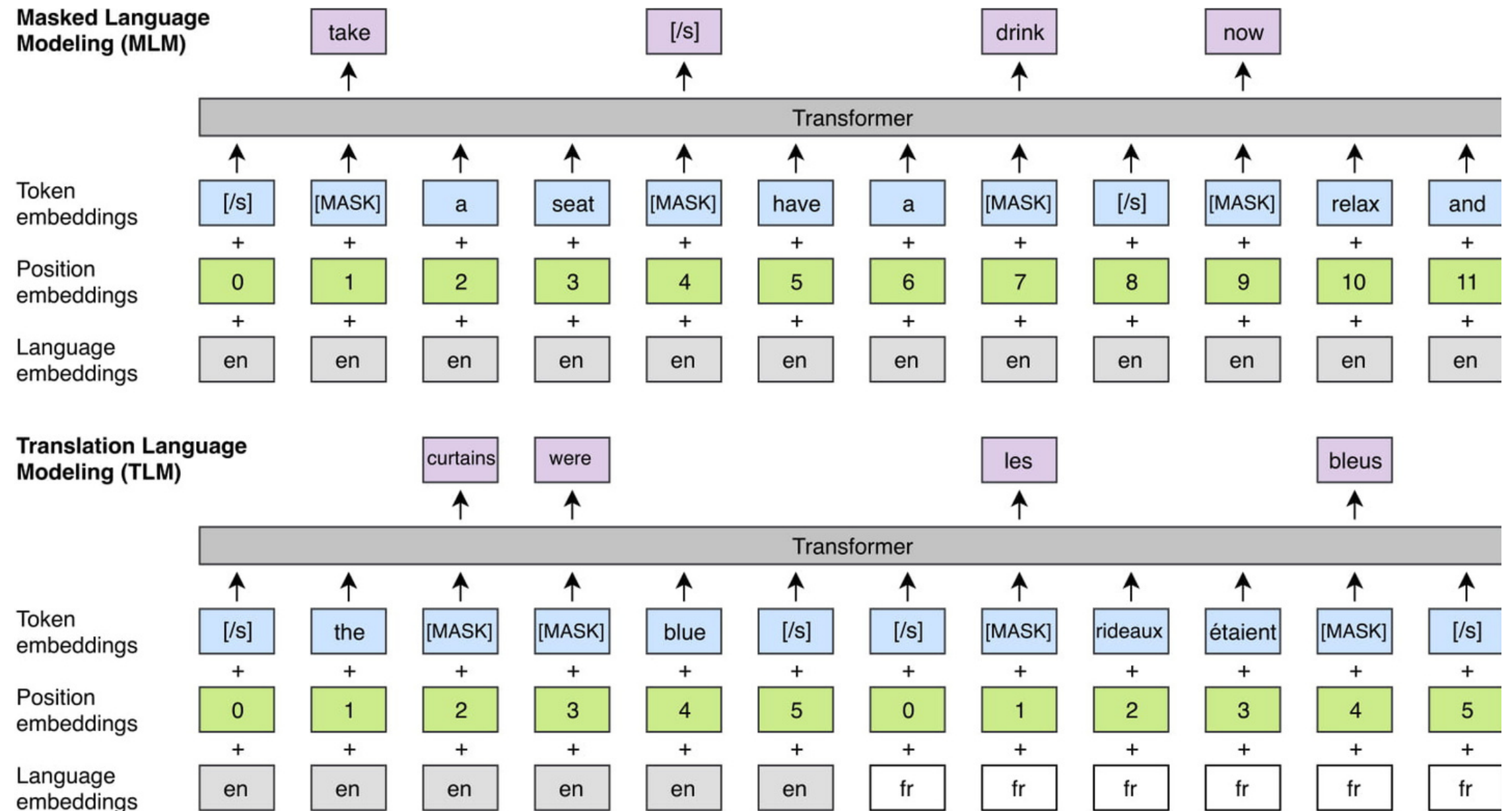**1** 3 pre-trained language modelling objectives are used: CLM, MLM, TLM.

2 The model is pre-trained over RoBERTa architecture.

3 Use the model for evaluation.

Back to Agenda Page

Masked Language Modeling (MLM)

	take			[/s]			drink		now	

Transformer

Token embeddings	[/s]	[MASK]	a	seat	[MASK]	have	a	[MASK]	[/s]	[MASK]	relax	and
Position embeddings	0	1	2	3	4	5	6	7	8	9	10	11
Language embeddings	en	en	en	en	en	en	en	en	en	en	en	en

Translation Language Modeling (TLM)

	curtains	were			les			bleus	

Transformer

Token embeddings	[/s]	the	[MASK]	[MASK]	blue	[/s]	[/s]	[MASK]	rideaux	étaient	[MASK]	[/s]
Position embeddings	0	1	2	3	4	5	0	1	2	3	4	5
Language embeddings	en	en	en	en	en	en	fr	fr	fr	fr	fr	fr

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	**99.92**	**82.04**	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
```
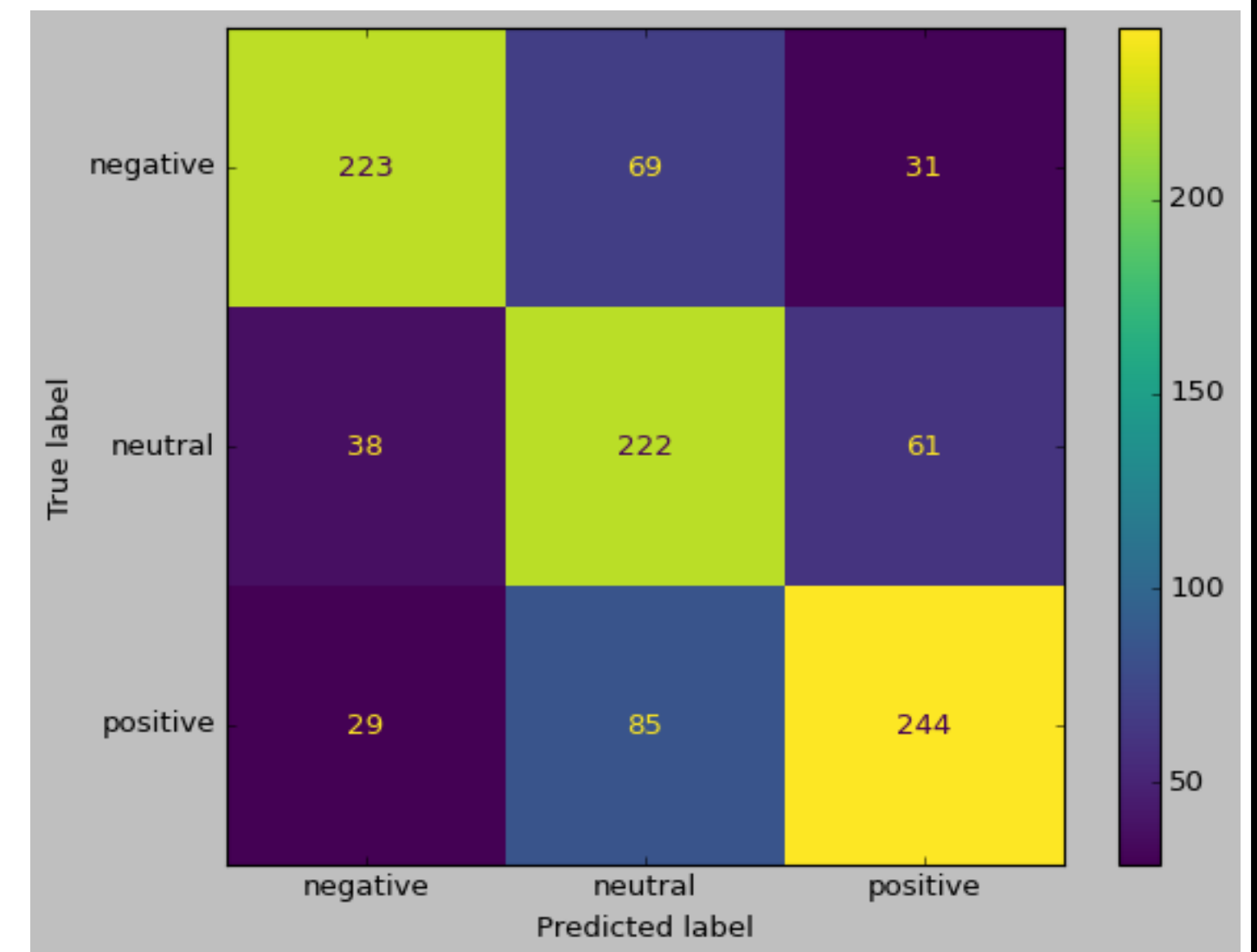
DIFFICULTIES AND LIMITATIONS
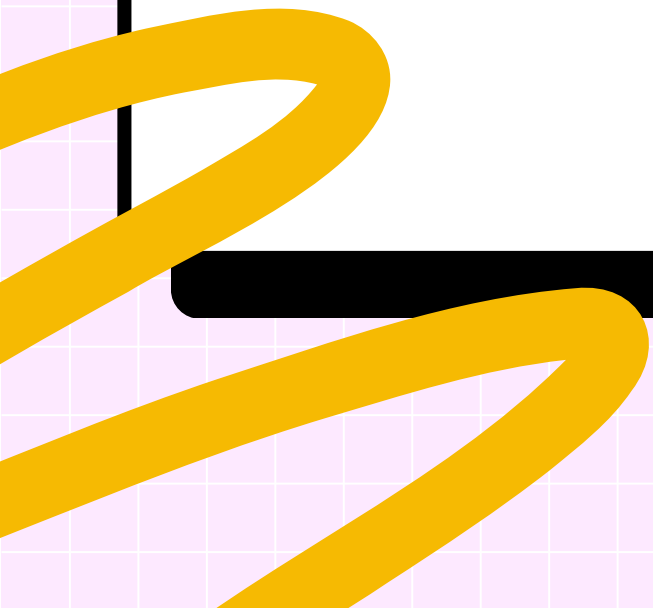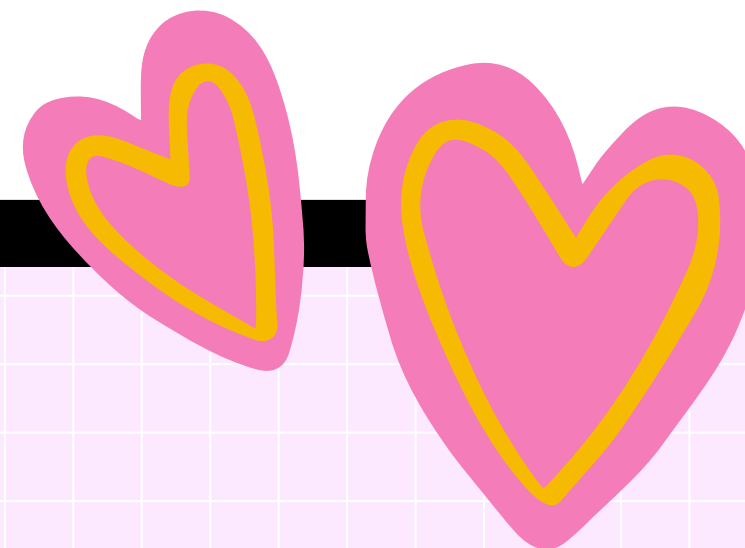
Difficulties

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
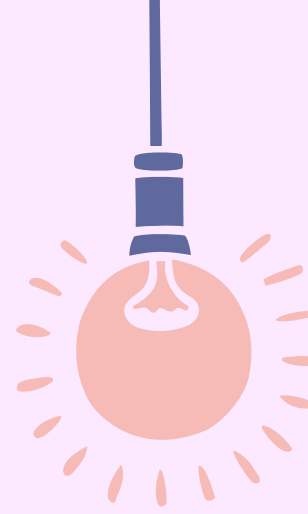
CONCLUSIONS

SUMMARY

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

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.