Twitter Sentiment Analysis

Indonesia's Presidential Election in 2019

Kaira Milani Fitria

27 September 2024

Background & Problem Statement

In the 2019 Indonesian election, millions of social media posts were shared, reflecting public opinion on various political topics. Understanding these sentiments can provide valuable insights into public opinion, which is critical for politicians and analysts. However, classifying these opinions accurately presents challenges due to the complexity and ambiguity of natural language. This project aims to tackle this challenge by analyzing public sentiments using AI models

Objectives

To compare the effectiveness of two machine learning models (Random Forest and LSTM) in classifying sentiment related to the twitter public comments about Indonesia Election in 2019

Scope

This analysis focuses on classifying sentiments (positive, neutral, and negative) from Twitter data. The models will be trained on text data after preprocessing, and their performance will be compared based on accuracy and other metrics.

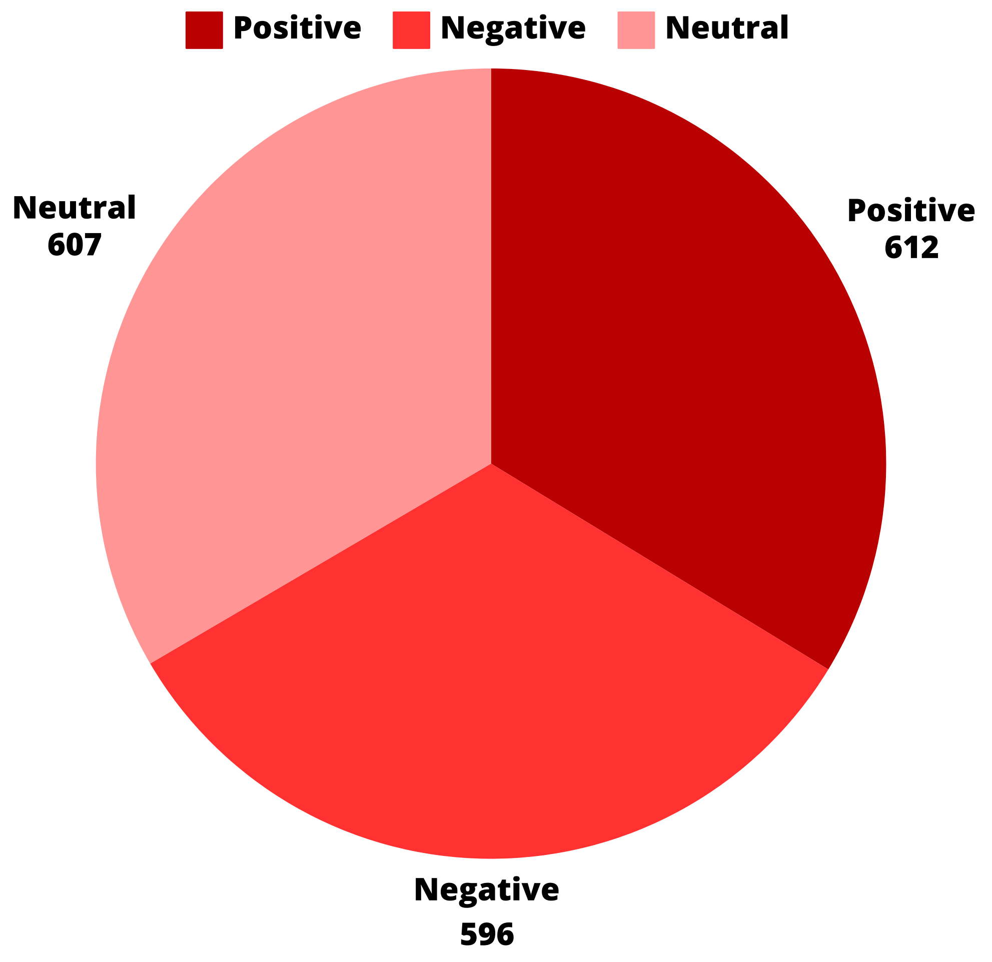
Data Collection

Dataset: Sentiment data from twitter users during the 2019 presidential election.

	sentimen	tweet
0	negatif	Kata @prabowo Indonesia tidak dihargai bangsa asing! Berita ini ðŸ’‡ pasti hoax buatan penguasa, ya kan @rockygerung?ðŸ˜œ https://twitter.com/mediaindonesia/status/1117575436337160192?s=21Â â€¦
1	netral	Batuan Langka, Tasbih Jokowi Hadiah dari Habib Luthfi Seharga Mercy? http://dlvr.it/R2pvZVÂ
2	netral	Di era Jokowi, ekonomi Indonesia semakin baik. #01IndonesiaMaju #JokowiLagi #JokowiMenangTotalDebat pic.twitter.com/W2ythnxsTp
3	positif	Bagi Sumatera Selatan, Asian Games berdampak pd ekonomi langsung diprediksi mencapai 18,5 triliun. Indonesia maju, Jokowi hebat!
4	negatif	Negara kita ngutang buat bngun infrastruktur yang udah dipake masyarakat, terus masyarakatnya ngeluh karena negara ngutang, setiap negara itu pasti ngutang, utang bisa dibayar kalo negara dapet penghasilan. Penghasilan negara itu ya dari pajak
...	...	...
1810	netral	Negarawan sejati sll bangga dan mengedepankan harga diri bangsanya yg berdaulat #2019GantiPresiden
1811	netral	1. HRS ceramah di Damai Indonesiaku 2. Perekonomian makin membaik. #PutihkanGBK
1812	netral	Mari bangun bangsa dgn mendukung perekonomian negara bersama Pak Jokowi. Ayo kerja! https://twitter.com/KaskusLoker/status/1111643312241295363Â â€¦
1813	netral	Bantu majukan perekonomian bangsa bersama Pak Jokowi, yuk! https://twitter.com/BKNSquare/status/1113655944955588610Â â€¦
1814	netral	Pak @jokowi mengubah cara pandang ekonomi. Kini semua orang terhubung, sehingga Indonesia menjadi lebih produktif dan efisien. #MenyatuPutihJokowi

1815 rows × 2 columns

Consists of 1815 tweet data containing three sentiment categories: Positive, Neutral and Negative.



Data Preprocessing

Data Cleaning

- Lowercase
- Remove URL
- Remove emojis
- Special characters
- Remove Punctuation
- Remove extra whitespace
- (Optional) remove stopwords

Additional Step :

- **Convert Slang Words**

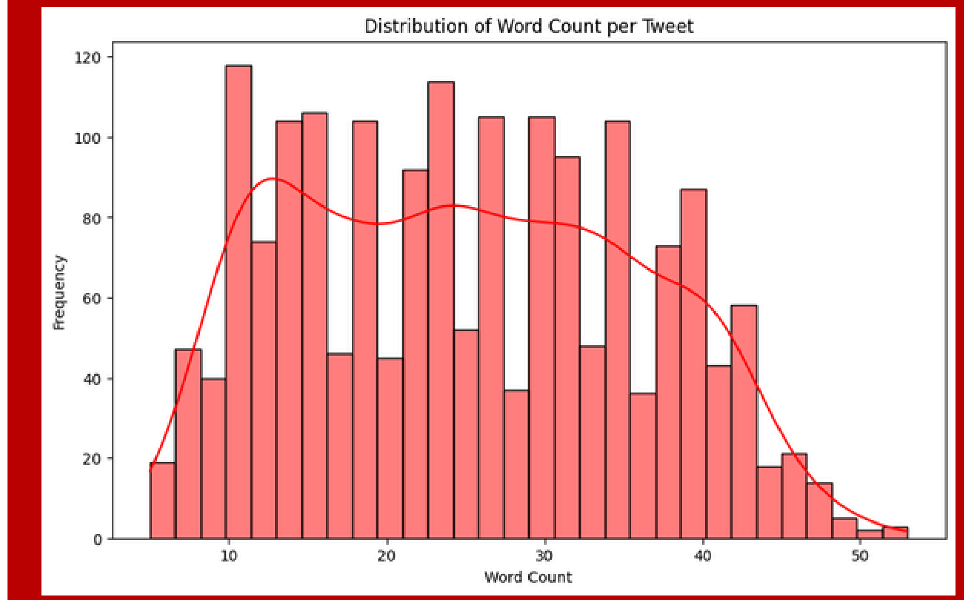
Hugging Face

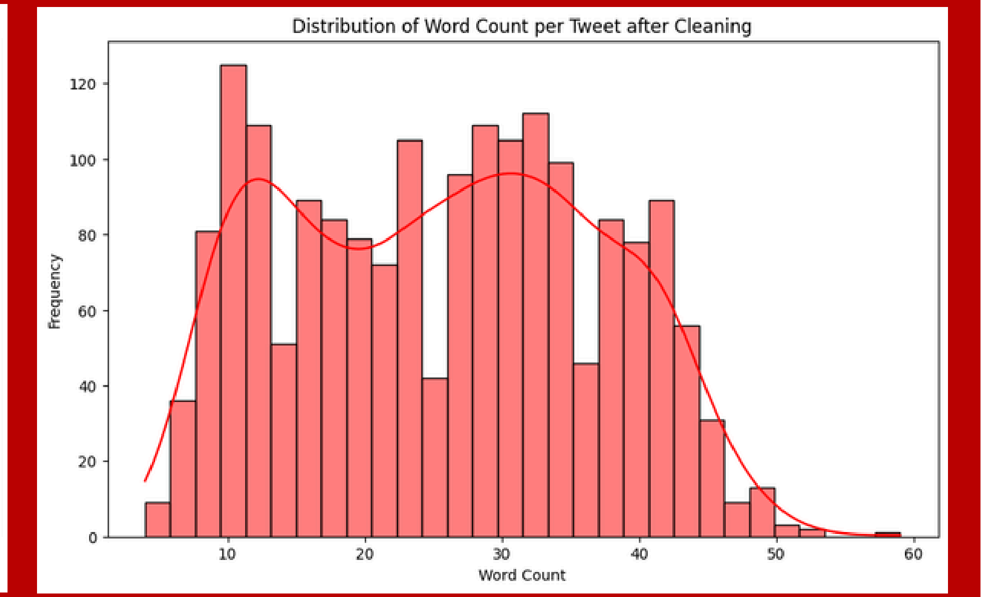
Datasets: theonlydo/**indonesia-slang**

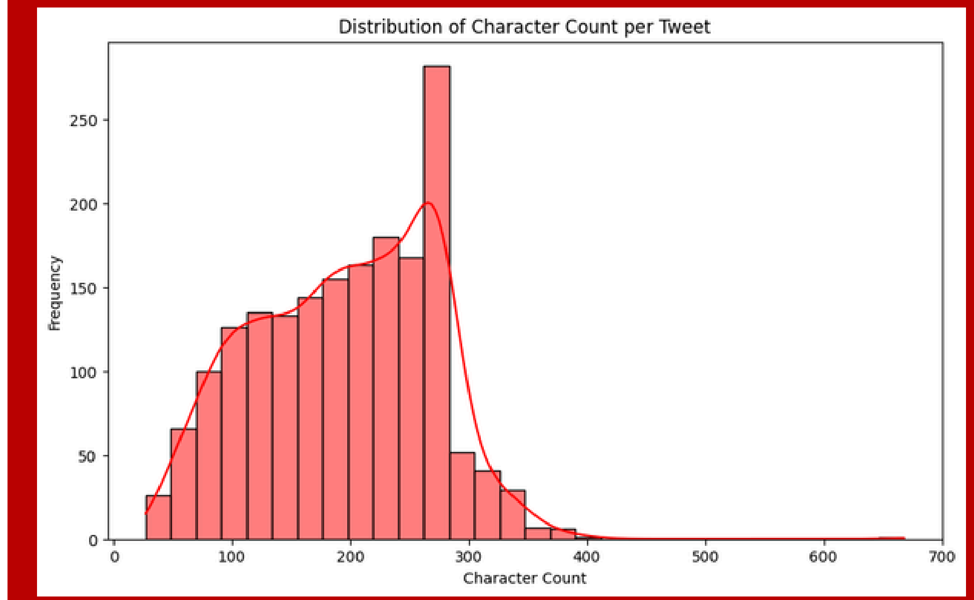
```
bngt → banget      lg → lagi
gk   → enggak      bkn → bukan
yg   → yang        ...    ...
```

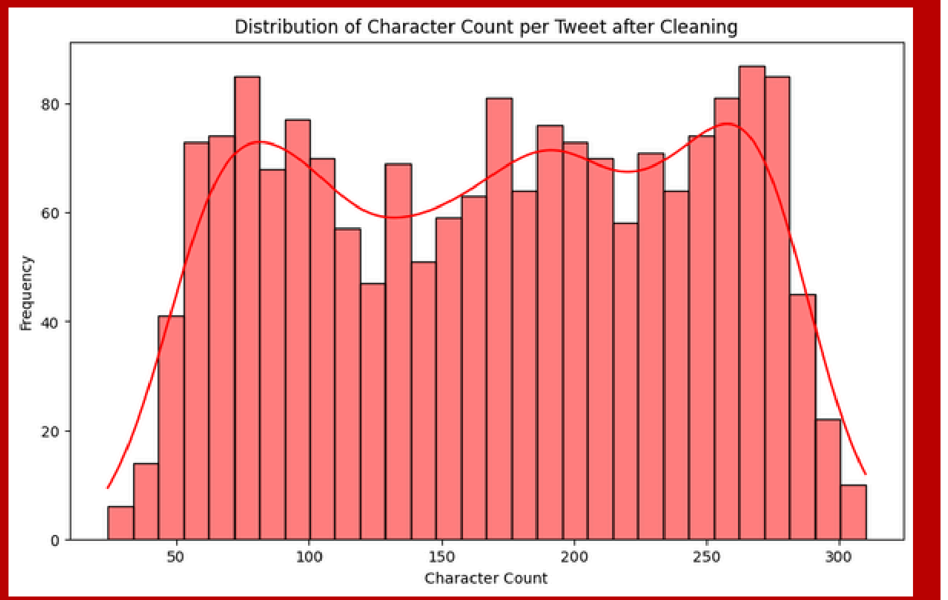
Comparison

Before Data Cleaning | **After Data Cleaning**





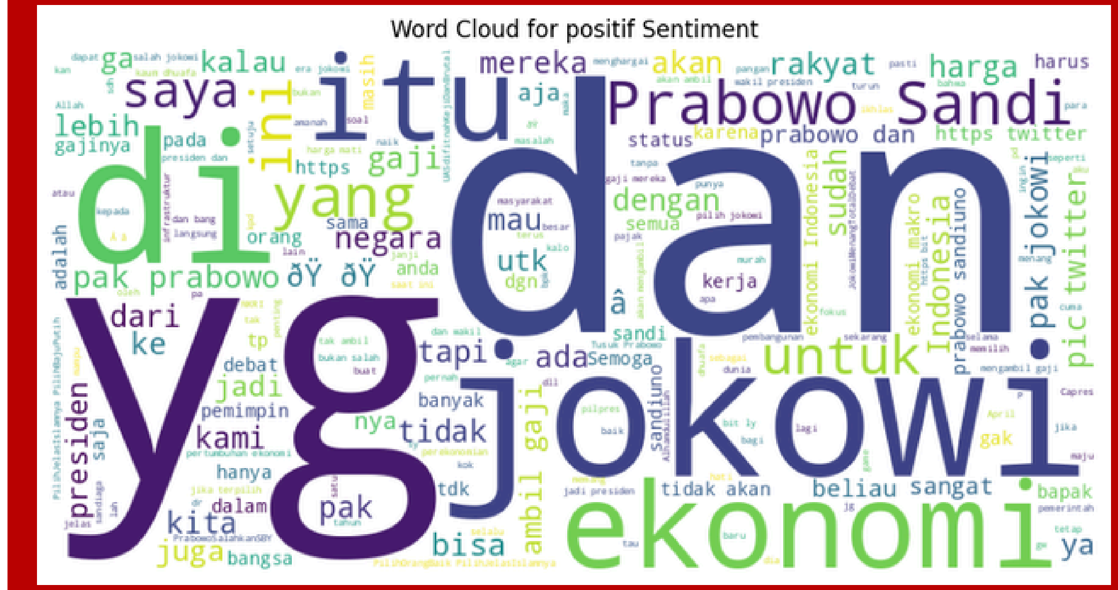


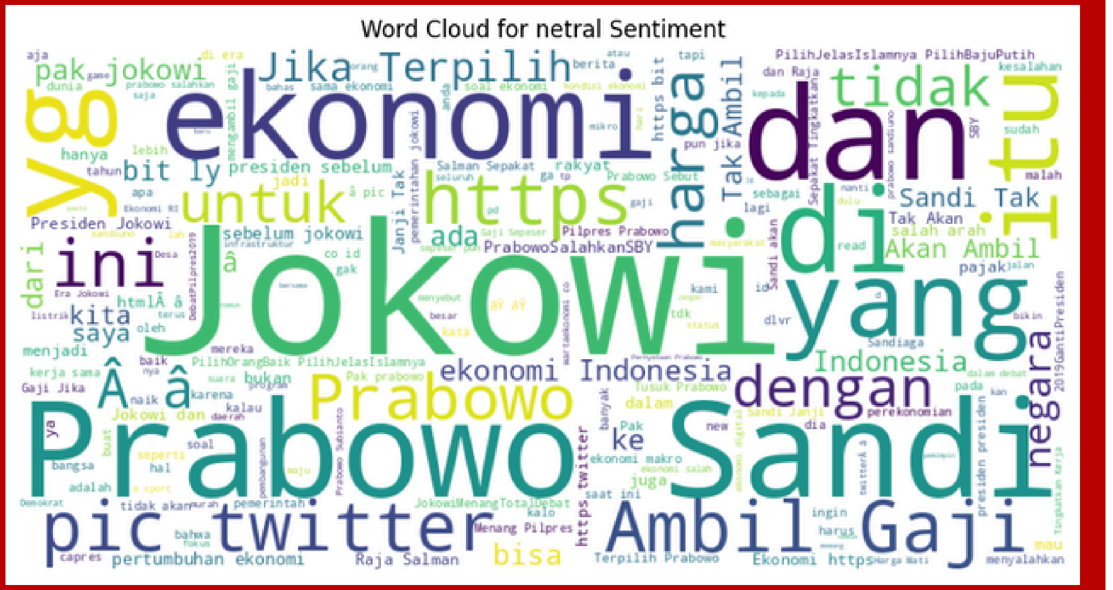


Data Preprocessing

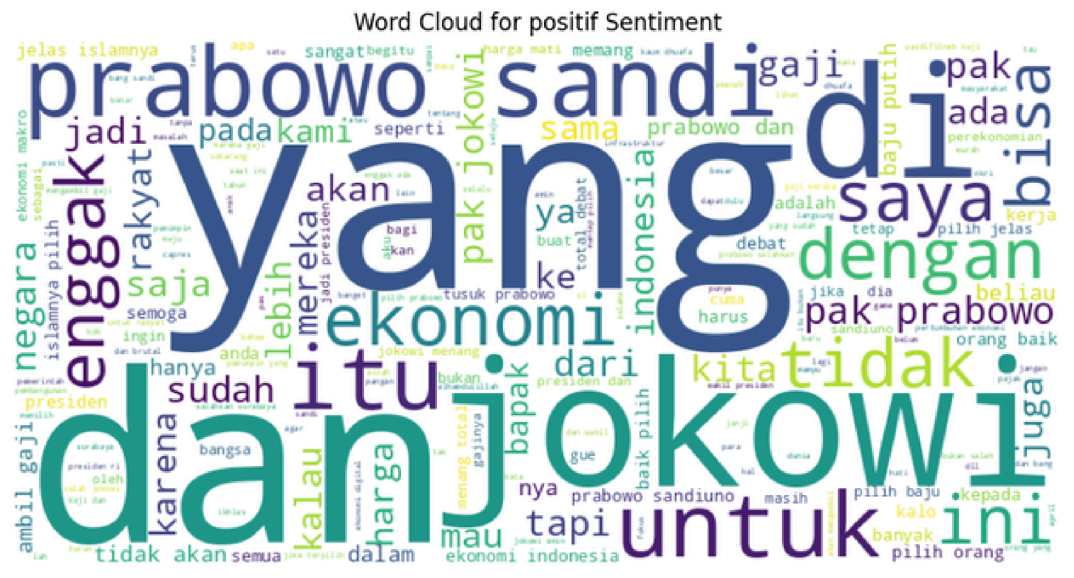
Word Cloud

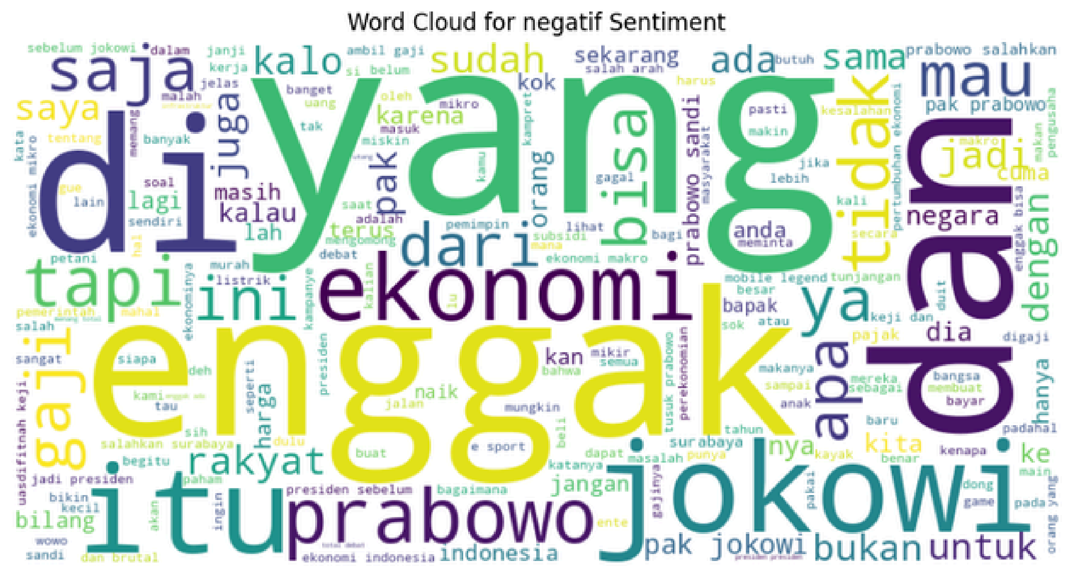
Before Data Cleaning

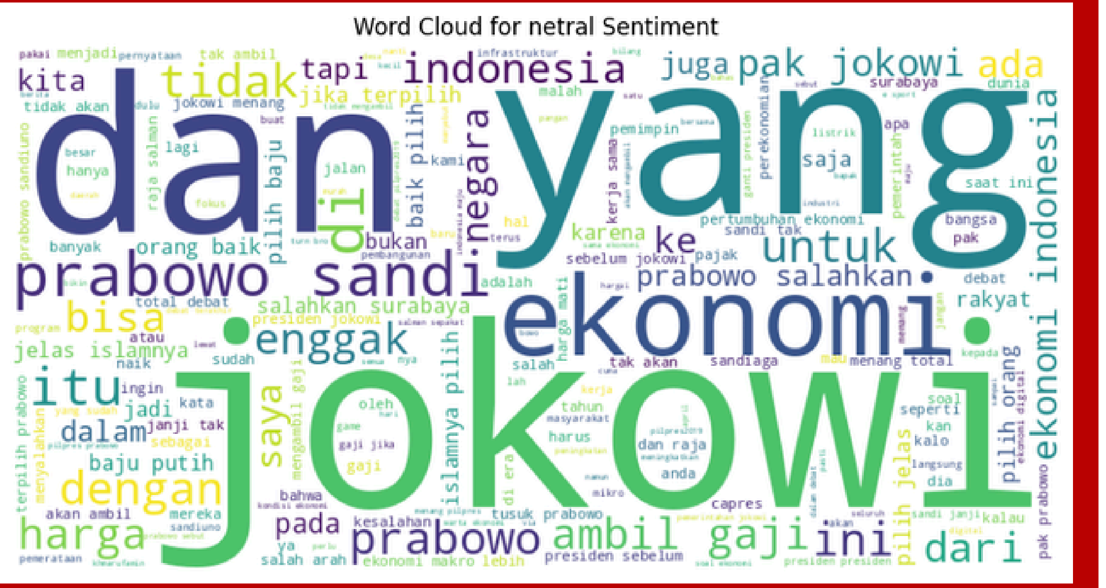




After Data Cleaning



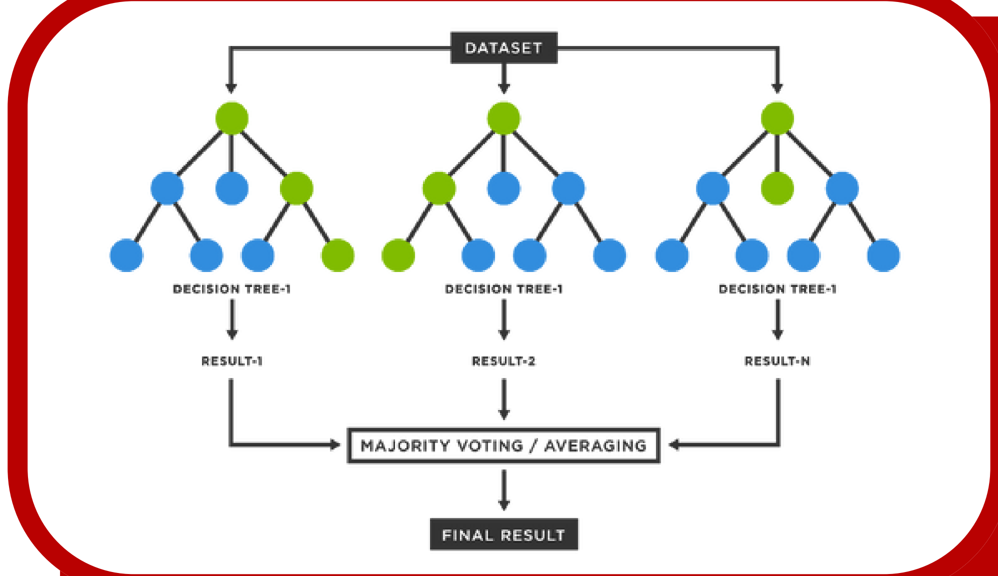




note : we are not using stopwords

Model Development

Comparison Architecture



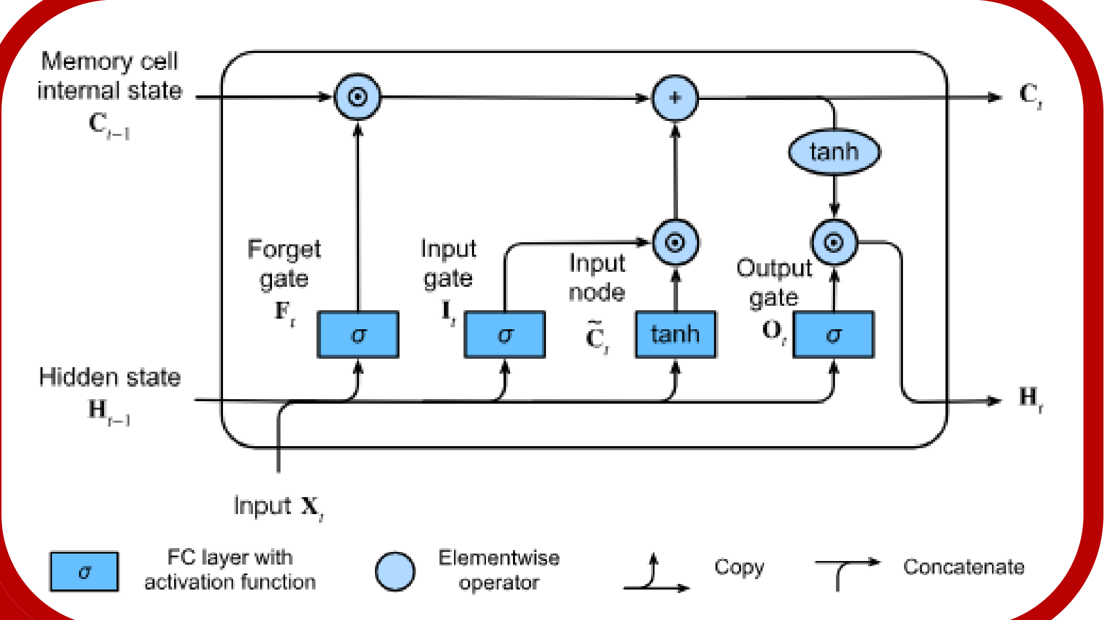
RANDOM FOREST

Random Forest uses multiple decision trees to make predictions based on majority vote.

+ Easy to understand and explain.
+ Works well with simple features like word counts.
+ Handles overfitting better than single trees.
- Can't understand word order or context.
- Might miss deeper patterns in language.

LSTM

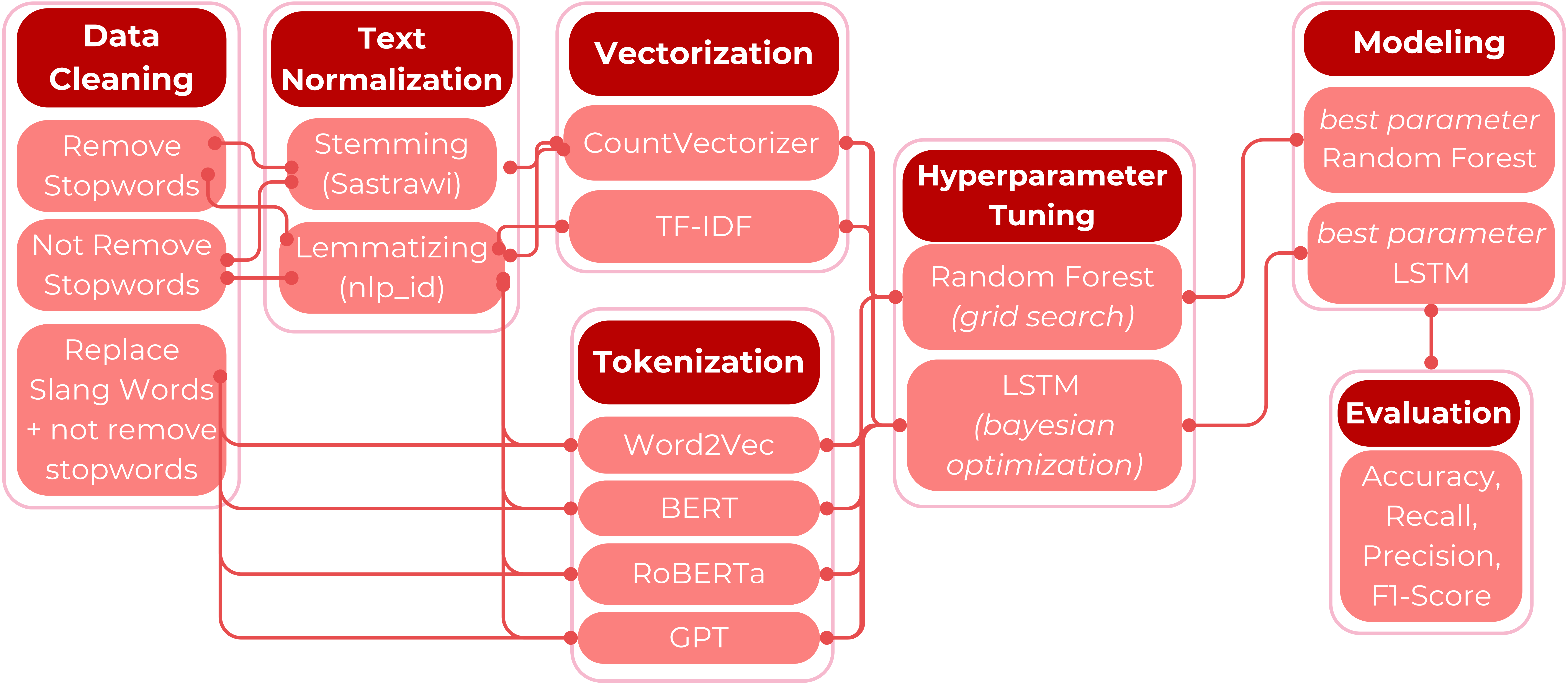
Long-Short Term Memory



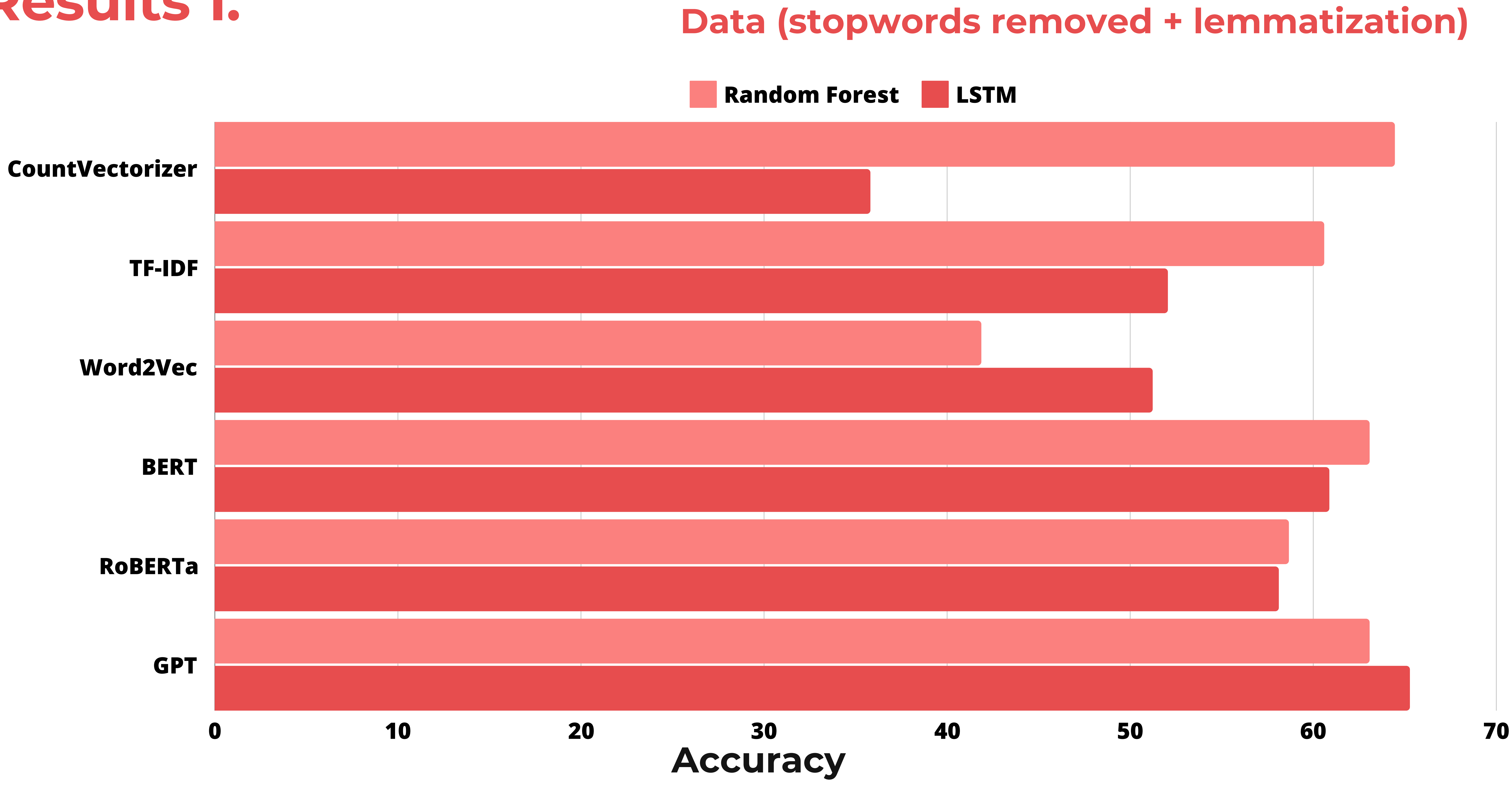
LSTM processes sequences of words by maintain memory of their order and relationships.

+ Captures word order and context
+ Can handle longer texts and more complex relationships between words.
+ Learns nuanced patterns, such as sarcasm or negation.
- Requires more data and resources to train.
- More complex to tune

Training & Optimization



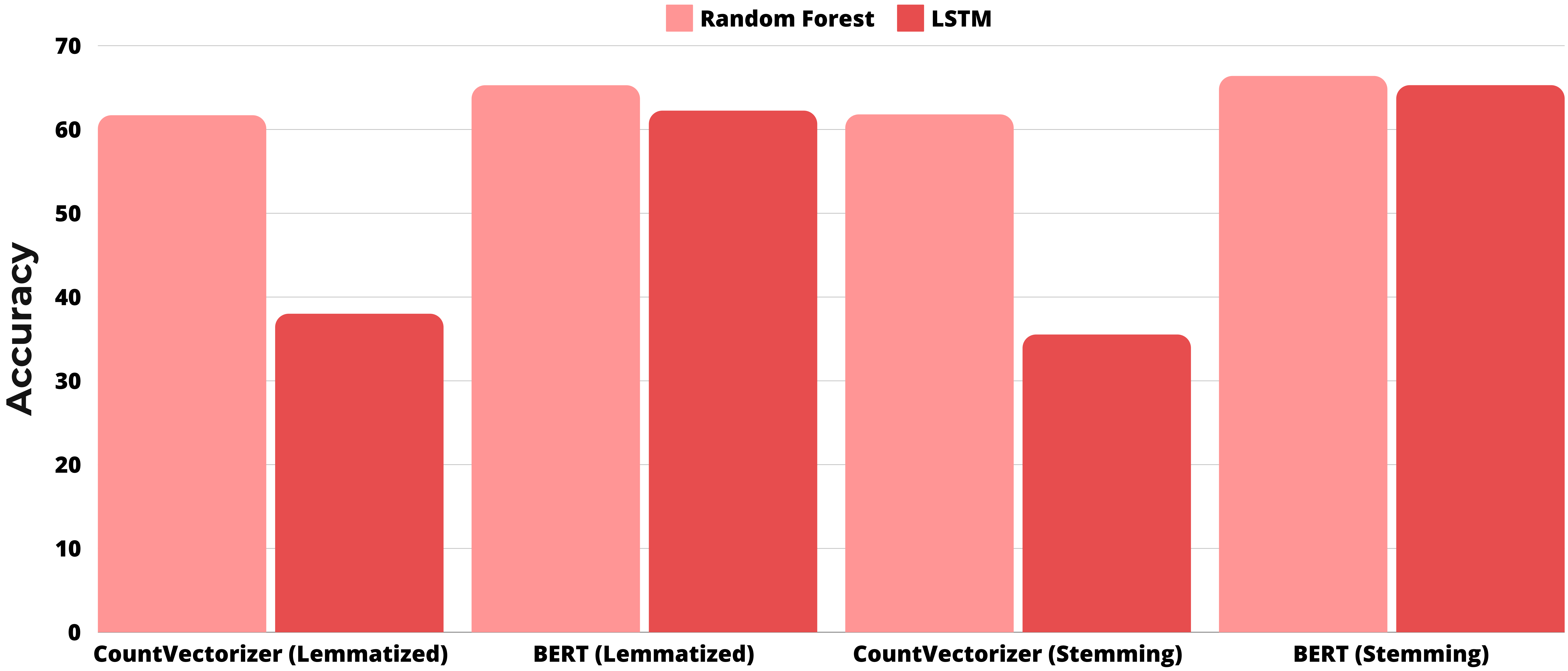
Results 1.



Data (stopwords removed + lemmatization)
Random Forest
LSTM
CountVectorizer
TF-IDF
Word2Vec
BERT
RoBERTa
GPT
0
10
20
30
40
50
60
70
Accuracy

Results 2.

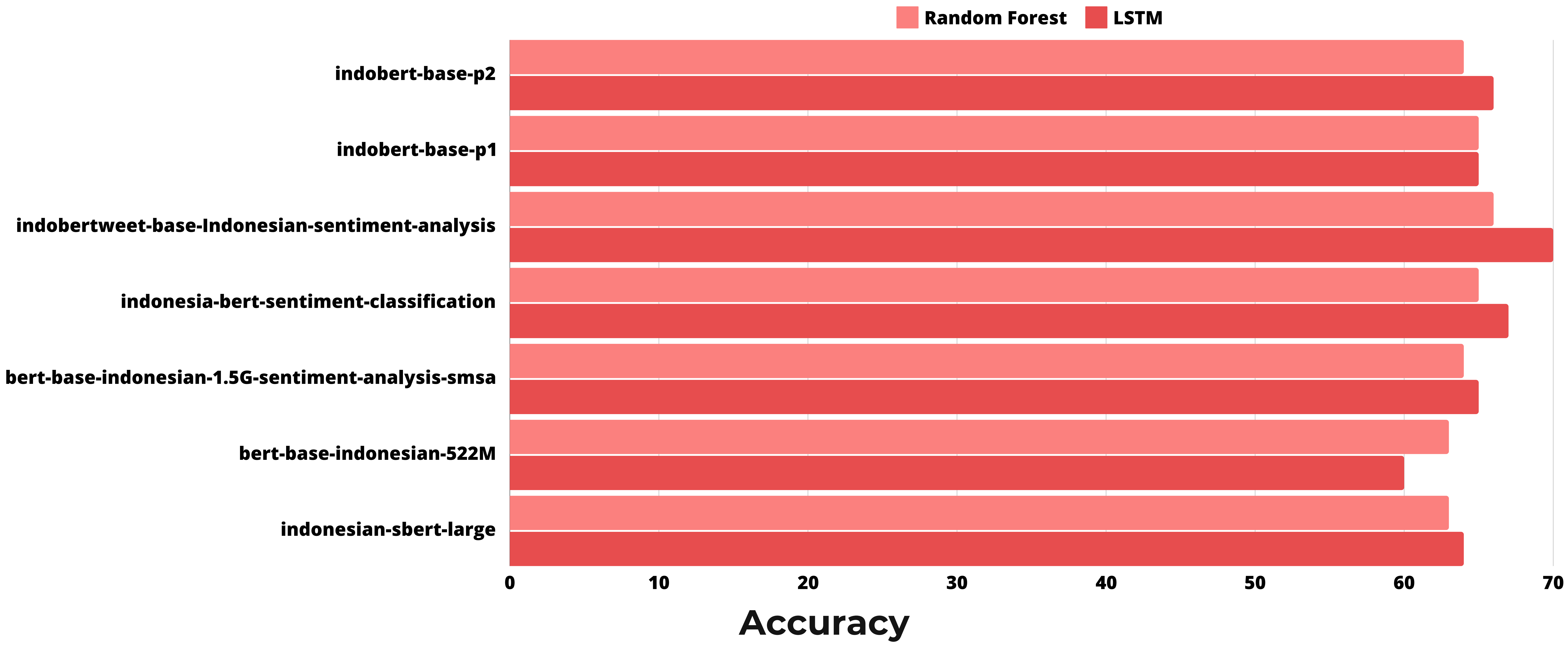
Data (stopwords + lemmatization/stemming)



Results 3.

Best Practice

Stopwords not removed + replace slang words + only tokenization embedding + hyperparameter tuning

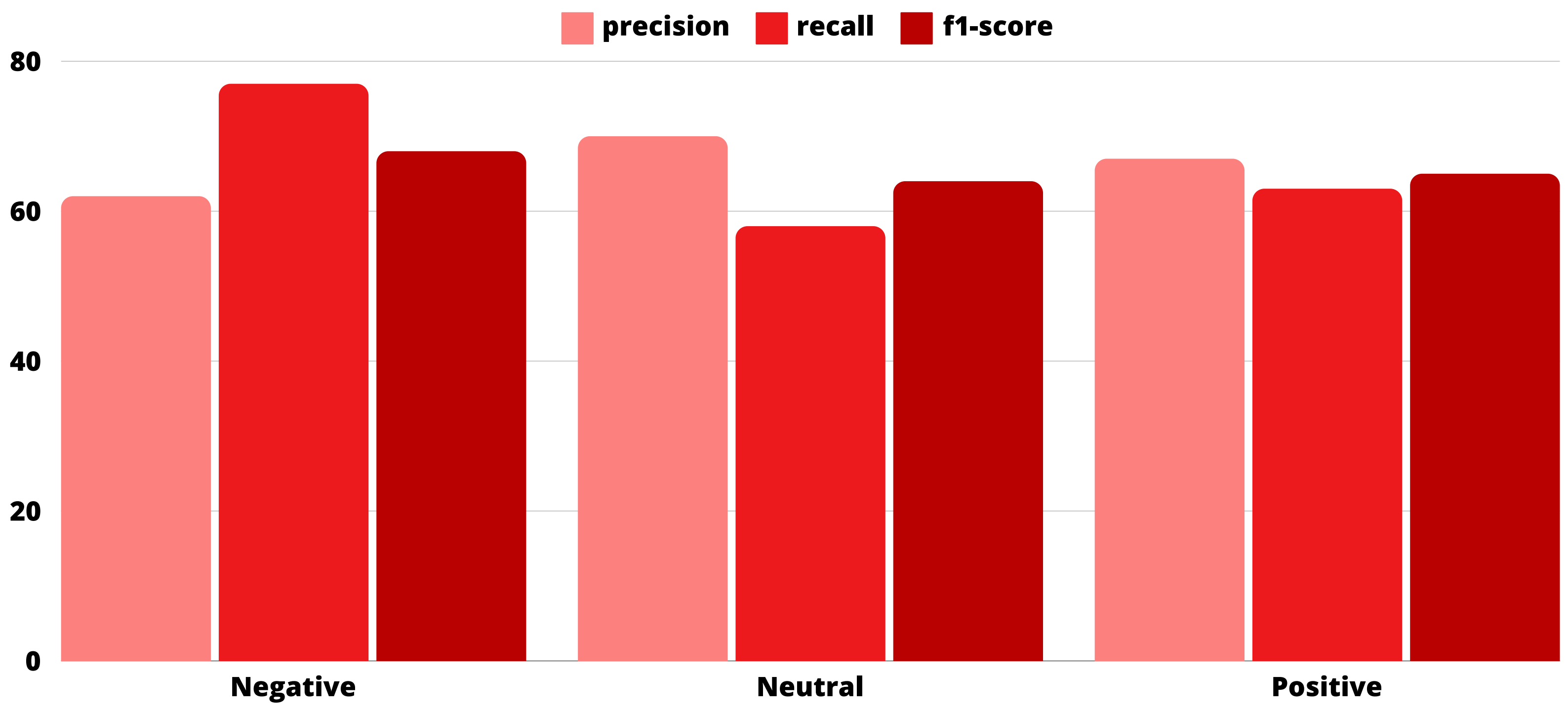


Results

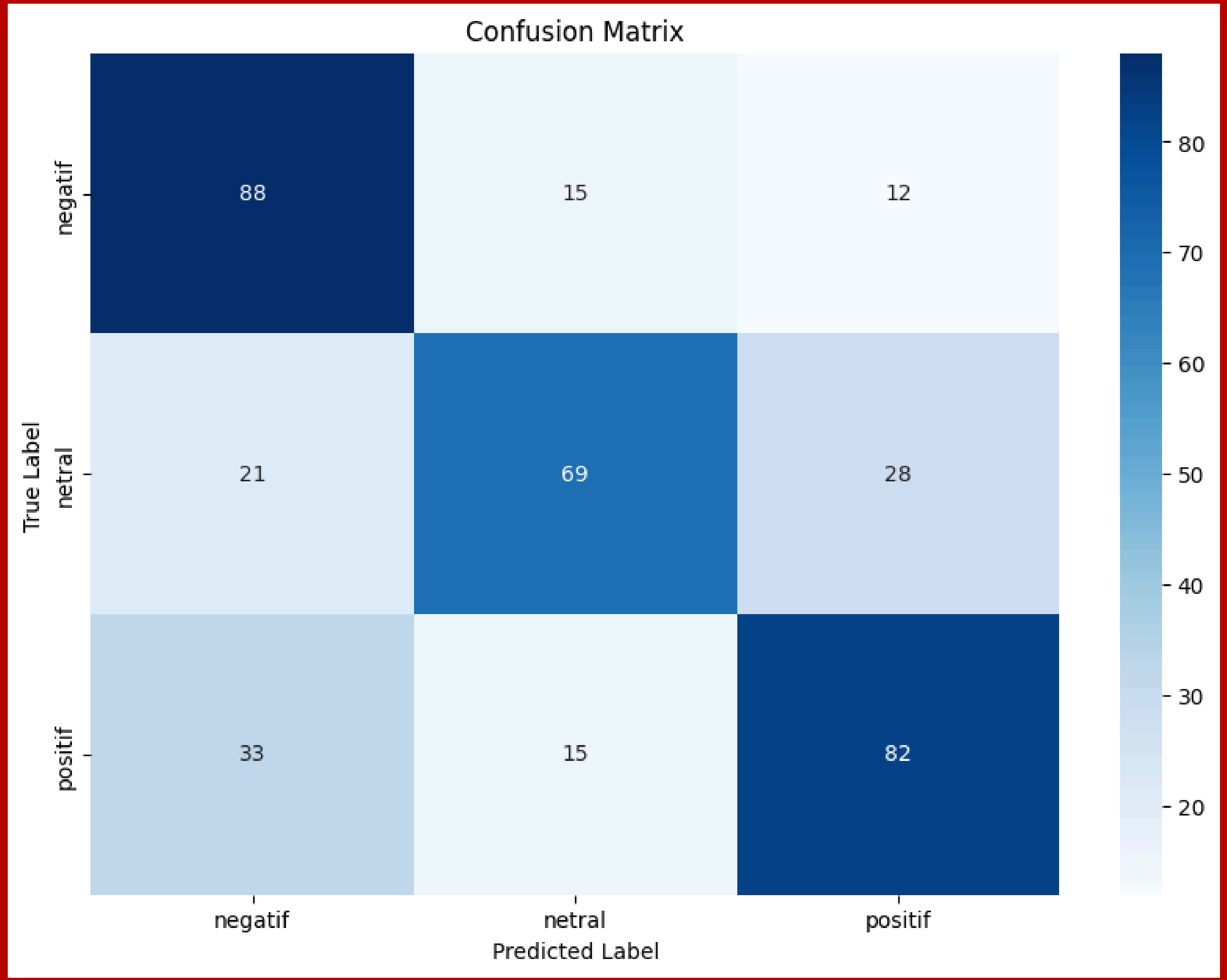
Best hyperparameter model based on test score

Random Forest

Parameter	Value
n_estimators	100
max_features	sqrt
criterion	gini
min_samples_split	2
min_samples_leaf	1
bootstrap	TRUE
random_state	42



Test Accuracy : 66%

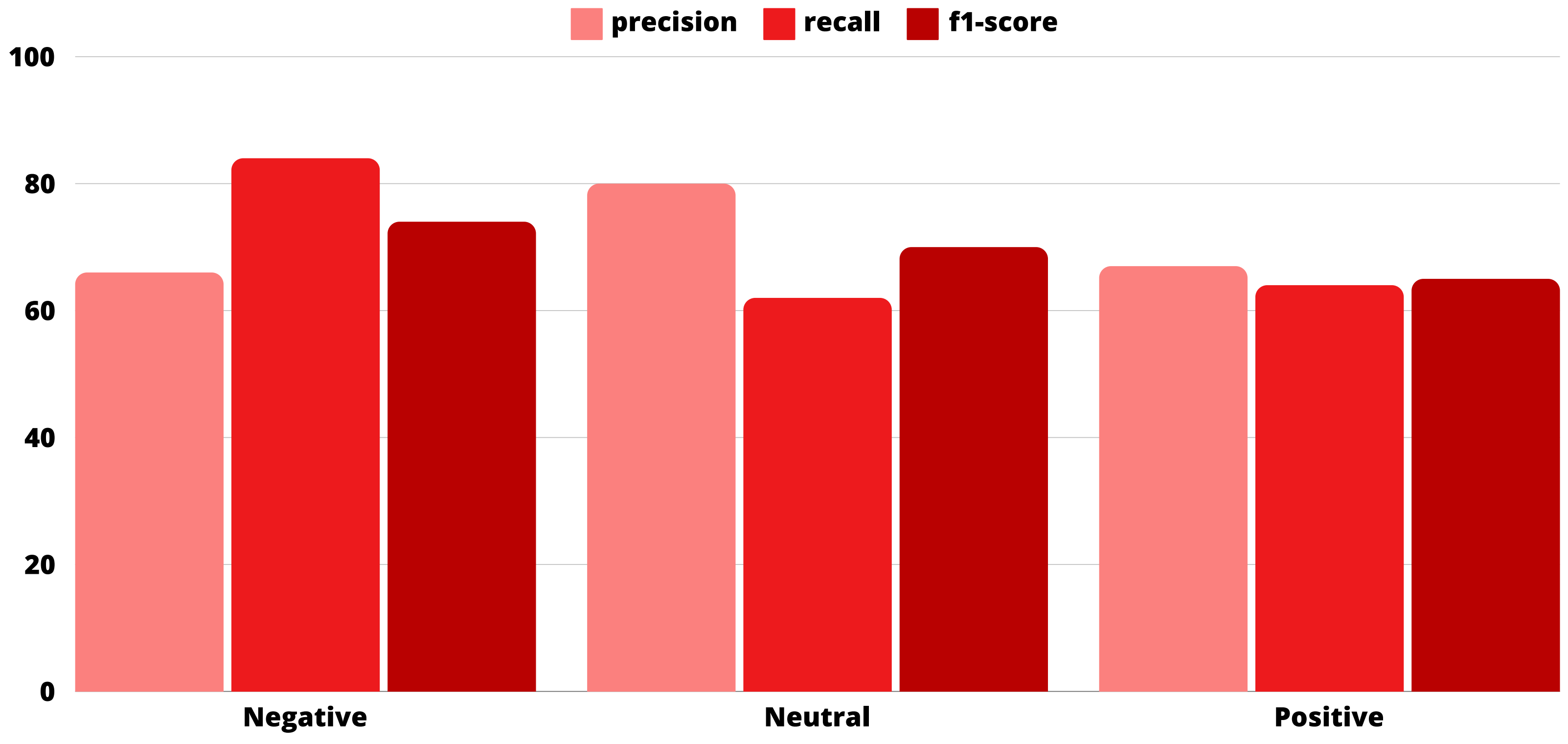


Results

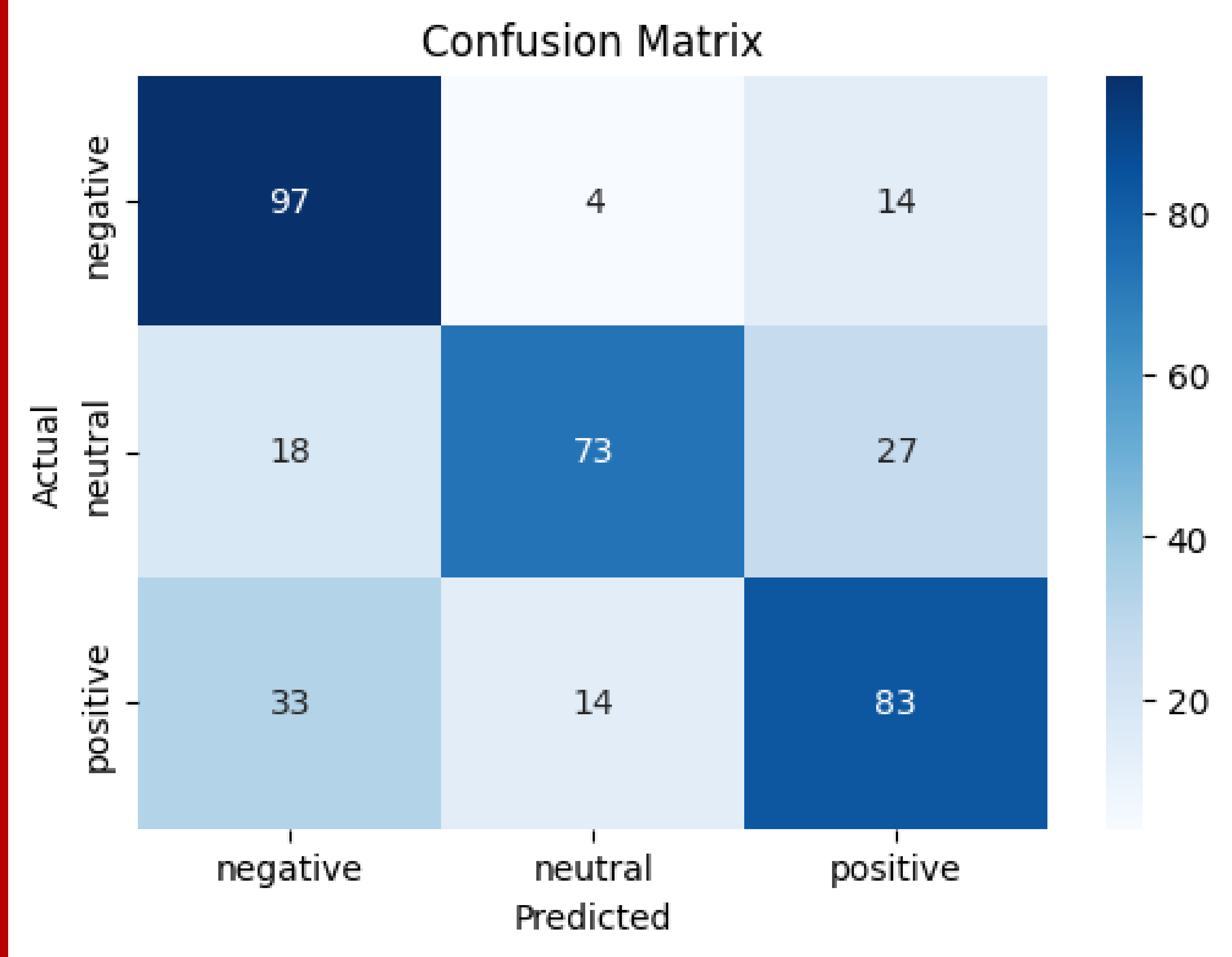
Best hyperparameter model based on test score

LSTM

Layer (type)	Shape	Param #
lstm_1 (LSTM)	(None, 64)	213,248
dropout_1 (Dropout)	(None, 64)	0
dense_2 (Dense)	(None, 128)	8,32
dense_3 (Dense)	(None, 3)	387
Optimizer		Adam



Test Accuracy : 70%



Future Improvement

Data Imbalance: Investigate methods to handle class imbalance, as more neutral or positive sentiments may be underrepresented.

Explore Advanced Embeddings: Experiment with more sophisticated embeddings like BERT, RoBERTa, or GPT to enhance LSTM performance by better capturing context and nuances in the text.

Optimize Preprocessing: Test different data cleaning approaches, such as selectively removing stopwords or handling slang, to improve the model's ability to understand context.

Conclusion

Random Forest works well with simple features like CountVectorizer or TF-IDF but struggles with understanding word order and context.

LSTM is more effective for sentiment analysis, using tokenizers like Word2Vec or BERT to capture word order and meaning.

Frequency-based vectorizers (CountVectorizer, TF-IDF) suitable for machine learning models like Random Forest, while embedding-based tokenizers work best with deep learning models like LSTM.

Data cleaning is key, but decisions on stopwords and slang replacement should depend on the context of the analysis.

Thank You

Kaira Milani Fitria

Kaira Milani Fitria

Documentations:

NLP-Projects/Project 1 Sentiment Analysis at main · kairamilanifitria/NLP...

Contribute to kairamilanifitria/NLP-Projects development by creating an account on GitHub.

GitHub

Result 1

RF

CountVectorizer

```
Accuracy: 0.6446280991735537
Classification Report:
              precision    recall  f1-score   support

    negative       0.63      0.74      0.68       115
     neutral       0.62      0.68      0.65       118
    positive       0.69      0.53      0.60       130

    accuracy                           0.64       363
   macro avg       0.65      0.65      0.64       363
weighted avg       0.65      0.64      0.64       363

Confusion Matrix:
 [[85 17 13]
 [20 80 18]
 [29 32 69]]
```

TF-IDF

```
Accuracy: 0.6060606060606061
Classification Report:
              precision    recall  f1-score   support

    negative       0.58      0.70      0.63       115
     neutral       0.59      0.60      0.60       118
    positive       0.66      0.53      0.59       130

    accuracy                           0.61       363
   macro avg       0.61      0.61      0.61       363
weighted avg       0.61      0.61      0.60       363

Confusion Matrix:
 [[80 21 14]
 [26 71 21]
 [33 28 69]]
```

Word2Vec

```
Accuracy: 0.418732782369146
Classification Report:
              precision    recall  f1-score   support

    negative       0.41      0.54      0.46       115
     neutral       0.42      0.34      0.38       118
    positive       0.43      0.38      0.41       130

    accuracy                           0.42       363
   macro avg       0.42      0.42      0.42       363
weighted avg       0.42      0.42      0.41       363

Confusion Matrix:
 [[62 23 30]
 [43 40 35]
 [48 32 50]]
```

BERT

```
Accuracy: 0.6308539944903582
Classification Report:
              precision    recall  f1-score   support

    negative       0.64      0.71      0.67       115
     neutral       0.62      0.61      0.62       118
    positive       0.64      0.58      0.60       130

    accuracy                           0.63       363
   macro avg       0.63      0.63      0.63       363
weighted avg       0.63      0.63      0.63       363

Confusion Matrix:
 [[82 19 14]
 [17 72 29]
 [30 25 75]]
```

RoBERTa

```
Accuracy: 0.5867768595041323
Classification Report:
              precision    recall  f1-score   support

    negative       0.58      0.64      0.61       115
     neutral       0.56      0.59      0.58       118
    positive       0.62      0.53      0.57       130

    accuracy                           0.59       363
   macro avg       0.59      0.59      0.59       363
weighted avg       0.59      0.59      0.59       363

Confusion Matrix:
 [[74 24 17]
 [23 70 25]
 [31 30 69]]
```

GPT

```
Accuracy: 0.6308539944903582
Classification Report:
              precision    recall  f1-score   support

    negative       0.60      0.70      0.65       115
     neutral       0.64      0.64      0.64       118
    positive       0.65      0.55      0.60       130

    accuracy                           0.63       363
   macro avg       0.63      0.63      0.63       363
weighted avg       0.63      0.63      0.63       363

Confusion Matrix:
 [[81 18 16]
 [20 76 22]
 [33 25 72]]
```

LSTM

CountVectorizer

```
Accuracy: 0.3581267217630854
Classification Report:
              precision    recall  f1-score   support

    negative       0.00      0.00      0.00       115
     neutral       0.00      0.00      0.00       118
    positive       0.36      1.00      0.53       130

    accuracy                           0.36       363
   macro avg       0.12      0.33      0.18       363
weighted avg       0.13      0.36      0.19       363

Confusion Matrix:
[[  0   1 114]
 [  0   0 118]
 [  0   0 130]]
```

TF-IDF

```
Test Accuracy: 0.5206611570247934
Classification Report:
              precision    recall  f1-score   support

    negative       0.45      0.86      0.59       115
     neutral       0.67      0.34      0.45       118
    positive       0.62      0.38      0.47       130

    accuracy                           0.52       363
   macro avg       0.58      0.53      0.50       363
weighted avg       0.58      0.52      0.50       363

Confusion Matrix:
 [[99  4 12]
 [59 40 19]
 [64 16 50]]
```

Word2Vec

```
Test Accuracy: 0.512396694214876
Classification Report:
              precision    recall  f1-score   support

    negative       0.61      0.38      0.47       115
     neutral       0.71      0.30      0.42       118
    positive       0.44      0.82      0.58       130

    accuracy                           0.51       363
   macro avg       0.59      0.50      0.49       363
weighted avg       0.58      0.51      0.49       363

Confusion Matrix:
 [[ 44   5  66]
 [ 14  35  69]
 [ 14   9 107]]
```

BERT

```
Test Accuracy: 0.6088154269972452
Classification Report:
              precision    recall  f1-score   support

    negative       0.65      0.57      0.61       115
     neutral       0.56      0.64      0.60       118
    positive       0.62      0.62      0.62       130

    accuracy                           0.61       363
   macro avg       0.61      0.61      0.61       363
weighted avg       0.61      0.61      0.61       363

Confusion Matrix:
 [[66 29 20]
 [14 75 29]
 [21 29 80]]
```

RoBERTa

```
Test Accuracy: 0.581267217630854
Classification Report:
              precision    recall  f1-score   support

    negative       0.57      0.60      0.58       115
     neutral       0.59      0.58      0.58       118
    positive       0.59      0.57      0.58       130

    accuracy                           0.58       363
   macro avg       0.58      0.58      0.58       363
weighted avg       0.58      0.58      0.58       363

Confusion Matrix:
 [[69 23 23]
 [21 68 29]
 [32 24 74]]
```

GPT

```
Test Accuracy: 0.6528925619834711
Classification Report:
              precision    recall  f1-score   support

    negative       0.66      0.70      0.68       115
     neutral       0.64      0.63      0.63       118
    positive       0.66      0.64      0.65       130

    accuracy                           0.65       363
   macro avg       0.65      0.65      0.65       363
weighted avg       0.65      0.65      0.65       363

Confusion Matrix:
 [[80 17 18]
 [20 74 24]
 [22 25 83]]
```

Result 2

RF

CountVectorizer

Stemming

```
Accuracy: 0.628099173553719
Classification Report:
              precision    recall  f1-score   support

    negative       0.61      0.70      0.65       115
     neutral       0.63      0.68      0.65       118
    positive       0.65      0.52      0.58       130

    accuracy                           0.63       363
   macro avg       0.63      0.63      0.63       363
weighted avg       0.63      0.63      0.63       363

Confusion Matrix:
 [[81 18 16]
 [18 80 20]
 [34 29 67]]
```

Lemmatized

```
Accuracy: 0.6170798898071626
Classification Report:
              precision    recall  f1-score   support

    negative       0.61      0.67      0.64       115
     neutral       0.60      0.69      0.64       118
    positive       0.64      0.51      0.57       130

    accuracy                           0.62       363
   macro avg       0.62      0.62      0.62       363
weighted avg       0.62      0.62      0.61       363

Confusion Matrix:
 [[77 23 15]
 [15 81 22]
 [34 30 66]]
```

BERT

Stemming

```
Accuracy: 0.6639118457300276
Classification Report:
              precision    recall  f1-score   support

    negative       0.64      0.75      0.69       115
     neutral       0.69      0.67      0.68       118
    positive       0.67      0.58      0.62       130

    accuracy                           0.66       363
   macro avg       0.67      0.67      0.66       363
weighted avg       0.67      0.66      0.66       363

Confusion Matrix:
 [[86 14 15]
 [16 79 23]
 [32 22 76]]
```

Lemmatized

```
Accuracy: 0.6528925619834711
Classification Report:
              precision    recall  f1-score   support

    negative       0.64      0.73      0.68       115
     neutral       0.67      0.66      0.67       118
    positive       0.65      0.58      0.61       130

    accuracy                           0.65       363
   macro avg       0.65      0.66      0.65       363
weighted avg       0.65      0.65      0.65       363

Confusion Matrix:
 [[84 15 16]
 [16 78 24]
 [32 23 75]]
```

LSTM

CountVectorizer

Stemming

```
Test Accuracy: 0.35537190082644626
Classification Report:
              precision    recall  f1-score   support

    negative       0.00      0.00      0.00       115
     neutral       0.00      0.00      0.00       118
    positive       0.36      0.99      0.53       130

    accuracy                           0.36       363
   macro avg       0.12      0.33      0.18       363
weighted avg       0.13      0.36      0.19       363

Confusion Matrix:
 [[  0   1 114]
 [  0   0 118]
 [  0   1 129]]
```

Lemmatized

```
Test Accuracy: 0.38016528925619836
Classification Report:
              precision    recall  f1-score   support

    negative       0.00      0.00      0.00       115
     neutral       0.39      0.57      0.46       118
    positive       0.38      0.55      0.45       130

    accuracy                           0.38       363
   macro avg       0.25      0.37      0.30       363
weighted avg       0.26      0.38      0.31       363

Confusion Matrix:
 [[ 0 49 66]
 [ 0 67 51]
 [ 1 58 71]]
```

BERT

Stemming

```
Test Accuracy: 0.6528925619834711
Classification Report:
              precision    recall  f1-score   support

    negative       0.68      0.63      0.65       115
     neutral       0.72      0.58      0.64       118
    positive       0.60      0.74      0.66       130

    accuracy                           0.65       363
   macro avg       0.66      0.65      0.65       363
weighted avg       0.66      0.65      0.65       363

Confusion Matrix:
 [[72 12 31]
 [15 69 34]
 [19 15 96]]
```

Lemmatized

```
Classification Report:
              precision    recall  f1-score   support

    negative       0.59      0.77      0.67       115
     neutral       0.70      0.50      0.58       118
    positive       0.60      0.61      0.61       130

    accuracy                           0.62       363
   macro avg       0.63      0.62      0.62       363
weighted avg       0.63      0.62      0.62       363

Confusion Matrix:
 [[88  9 18]
 [25 59 34]
 [35 16 79]]
```

Best Result

RF

Random Forest Classification Report:

	precision	recall	f1-score	support
0	0.62	0.77	0.68	115
1	0.70	0.58	0.64	118
2	0.67	0.63	0.65	130
accuracy			0.66	363
macro avg	0.66	0.66	0.66	363
weighted avg	0.66	0.66	0.66	363

indobenchmark/indobert-base-p1

LSTM

Classification Report:

	precision	recall	f1-score	support
negative	0.66	0.84	0.74	115
neutral	0.80	0.62	0.70	118
positive	0.67	0.64	0.65	130
accuracy			0.70	363
macro avg	0.71	0.70	0.70	363
weighted avg	0.71	0.70	0.69	363

Aardiiiiy/indobertweet-base-Indonesian-sentiment-analysis