The models produce better accuracy WITHOUT preprocessing at all.

It's probably because Marbert is trained on raw data so when we feed it raw data it gives a better embedding.

There definitely is incorrectly labeled data.

There is ambiguous data that is hard to determine wither the True label is correct or not

Classes that are often confused together:

1. Love and joy
2. Sadness and anger
3. Sadness and sympathy
4. None and most of them

In case of deliberate overfitting of the (obviously larger than necessary) model rarely reaches over 90% accuracy and the best case is 95% accuracy.

In a normal model a case of overfitting does not even reach 90% accuracy.

Cleaned data accuracy:

A table of numbers with black text

Description automatically generated

Light stemming accuracy:

A table of numbers with black text

Description automatically generated

Raw data accuracy: (no emojis)

A table of numbers with black text

Description automatically generated

Raw data accuracy:

A screenshot of a table

Description automatically generated

A chart with different colored squares

Description automatically generated

Some examples can’t be labeled as one of the eight labels.