**Laporan Kesimpulan**

**Klasifikasi *Multilabel Dangerous Speech Twitter* menggunakan *Pre-trained Model* IndoBERTweet**

**Link Colab:**

<https://colab.research.google.com/drive/1F7bVwfiVAcatYliueqIHsfaThNVuHPIc>

**Langkah-langkah yang dilakukan:**

1. Melakukan *installing* dan *importing* *library*

Library yang digunakan seperti: wordcloud, scikit-learn, pandas dan lainnya.

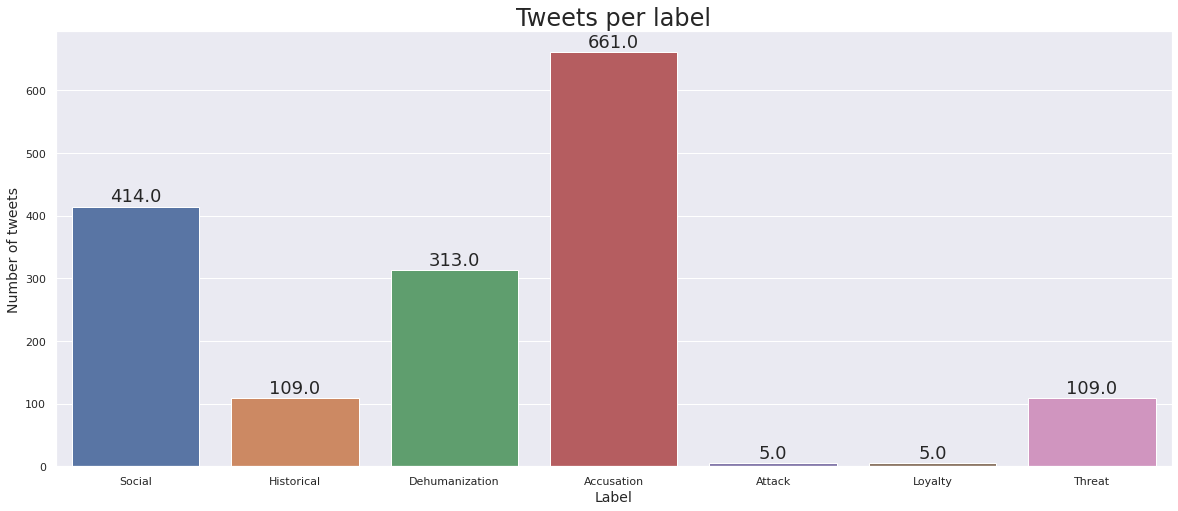
1. Melakukan *importing* data dari Google Drive

Data yang digunakan dapat diakses di:

[​​https://drive.google.com/file/d/1EA0\_fV2Z\_PsdNM4fdQsKTjbffEWWDZum/view?usp=share\_link](https://drive.google.com/file/d/1EA0_fV2Z_PsdNM4fdQsKTjbffEWWDZum/view?usp=share_link)

1. Eksplorasi data

Pada tahap ini, dilakukan pemahaman terhadap data yang digunakan. Dari hasil eksplorasi data yang dilakukan, didapatkan jumlah total data ada 851 data *tweet* yang bisa digunakan dengan sebaran kategori label sebagai berikut:



**\****Social* = *Social Context*

*Historical* = *Historical Context*

*Dehumanization* = *Dehumanization*

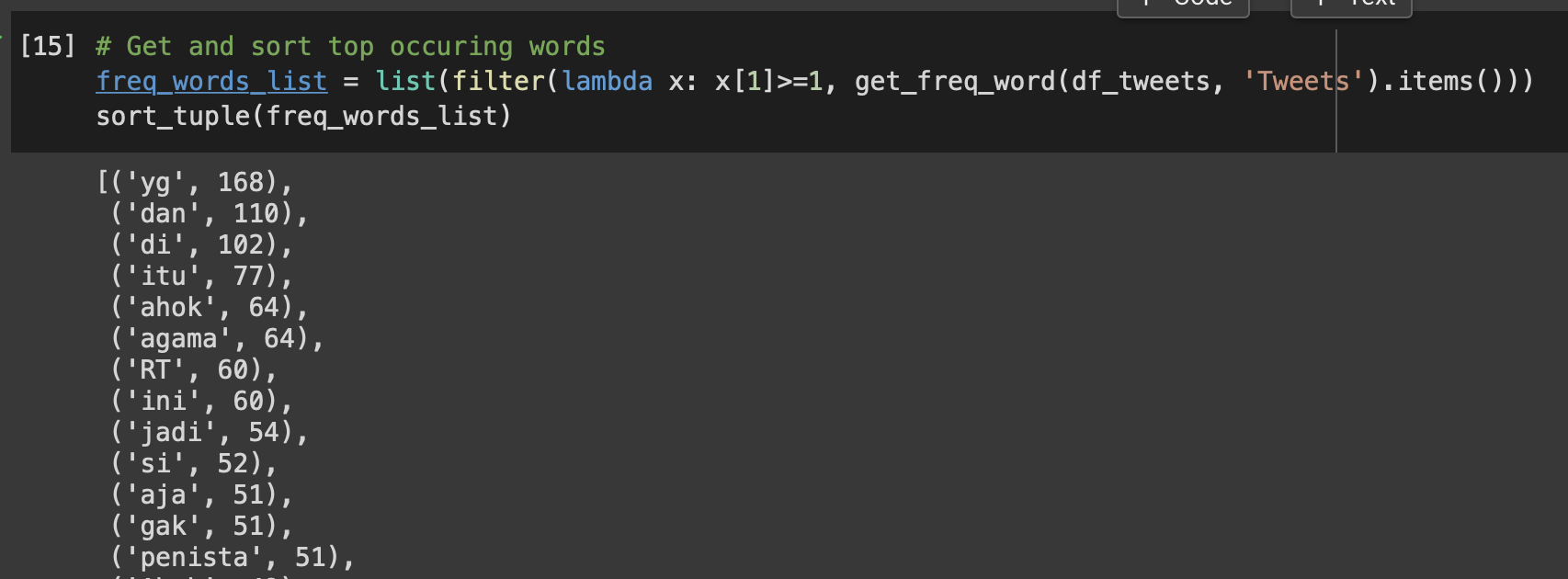
*Accusation* = *Accusation in a mirror*

*Attack* = *Assertion of Attack against Women and Girls*

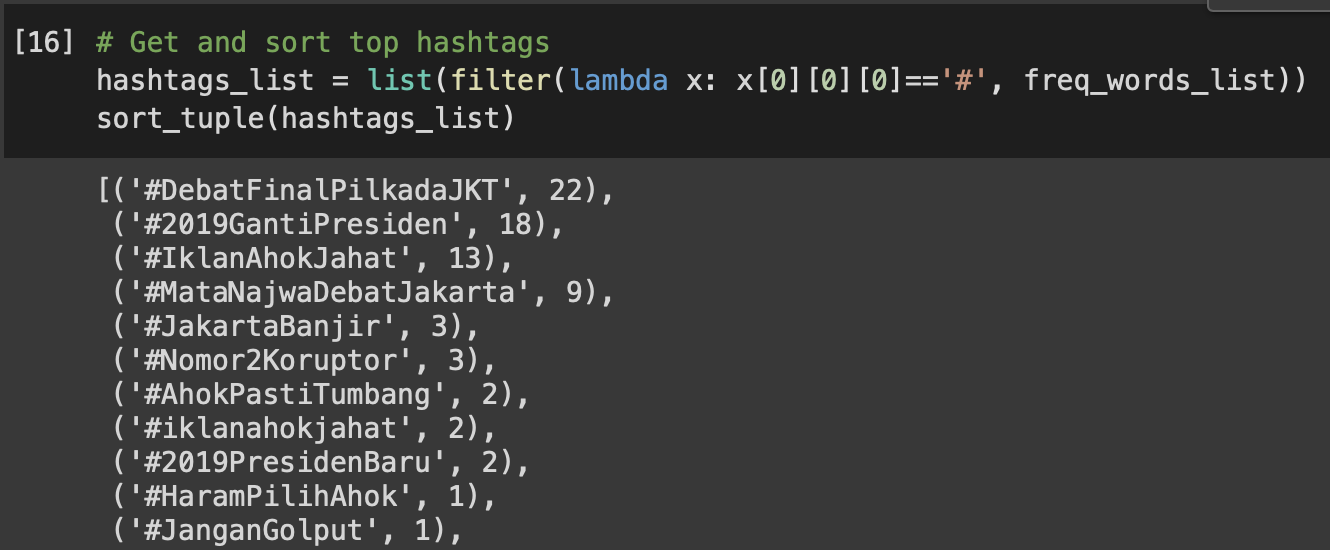
*Loyalty* = *Questioning In-group Loyalty*

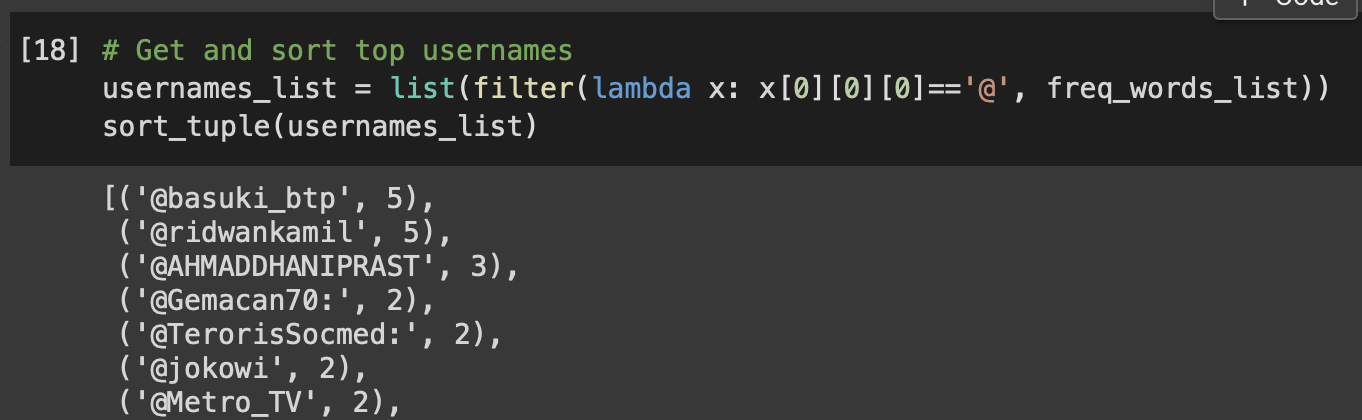
*Threat* = *Threat to Group Integrity or Purity*

Selain itu, juga ditambahkan fungsi untuk menunjukkan kata-kata yang paling sering muncul dari keseluruhan data, ini dapat digunakan untuk *stopword removal*, namun karena *stopword* yang digunakan hanya satu kata, yaitu ‘rt’, bagian ini bisa diabaikan.

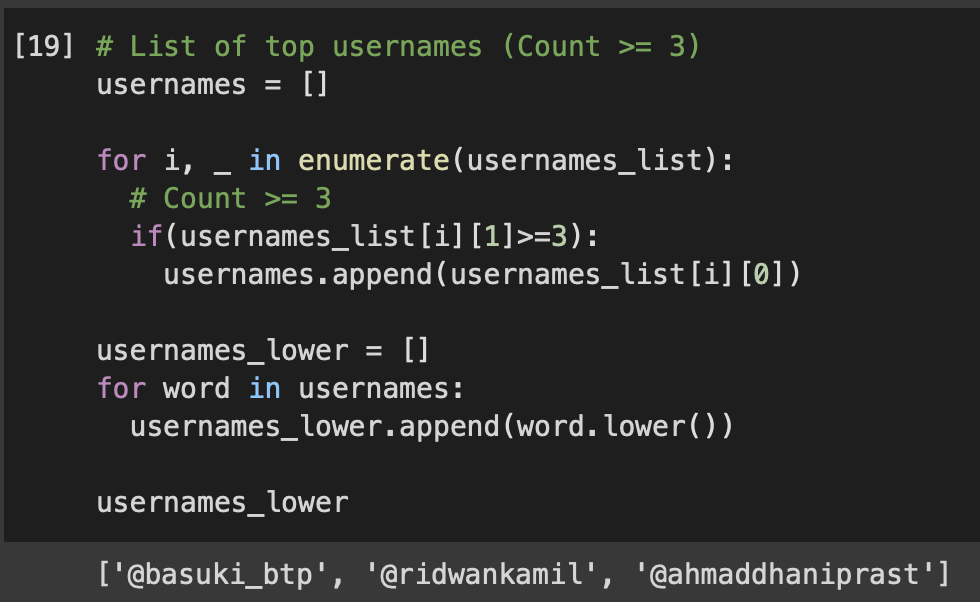
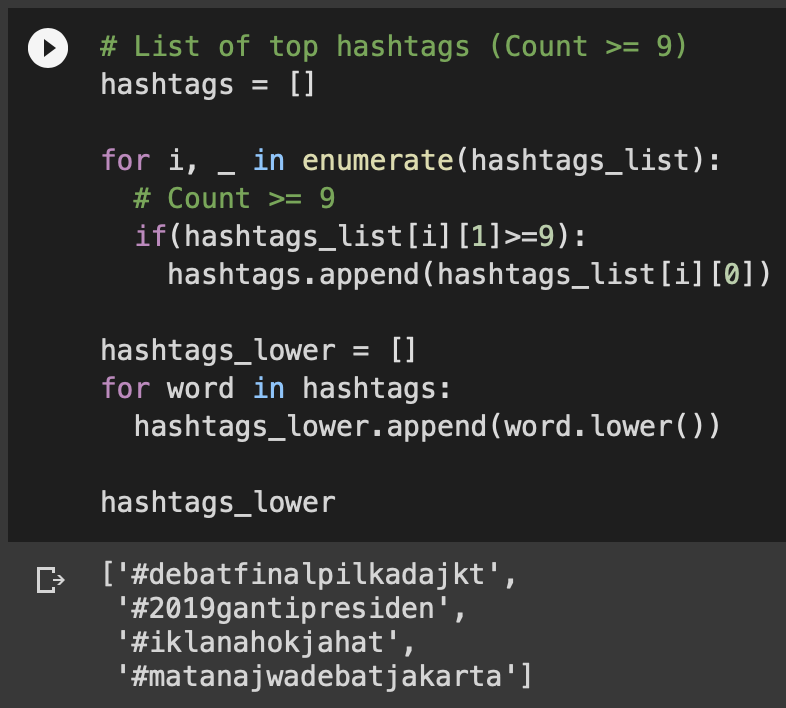


Selain itu, juga ditambahkan fungsi untuk menunjukkan *hashtag* dan *username* yang paling sering di-*tweet* dan di-*mention*, hal ini ditambahkan sebagai salah satu skenario perbandingan, yaitu data yang diproses dengan menambahkan *hashtag* dan *username* yang paling sering muncul, atau dengan mengabaikan *hashtag* dan *username* itu sendiri.





Berikut merupakan hasil hashtag dan username yang paling sering muncul:



Setelah menambahkan berbagai fungsi tadi, dibuat sebuah WordCloud dari data yang dipakai, berikut merupakan WordCloud per label untuk data yang belum diproses atau masih mentah:



1. Pra-proses data

Pada tahap ini, dilakukan pra-proses data *tweet* agar bisa didapatkan data *tweet* bersihnya. Tahap-tahap pra-proses adalah sebagai berikut:

1. *Case* *folding*
2. Mengubah entitas HTML jika memang ada (‘&amp;’, ‘&lt;’, ‘&gt;’)
3. Mengubah hashtag dan username yang sering muncul menjadi bentuk panjang dan aslinya (contoh: ‘#debatfinalpilkadajkt’ menjadi ‘debat final pilkada jakarta’, ‘@basukibtp’ menjadi ‘ahok’, dan seterusnya)
4. Menghapus awalan *link*
5. Menghapus HTML *tag*
6. Menghapus tanda *hashtag* pada *hashtag* yang tidak sering muncul
7. Mengubah *mention* *user* yang tidak sering muncul menjadi ‘@USER’
8. Menghapus karakter yang bukan huruf
9. Menghapus spasi yang berlebihan
10. Melakukan *stopwords removal*, pada kasus ini stopwords yang digunakan hanya satu kata, yaitu ‘rt’.
11. Melakukan *lemmatizing*
12. Membuang data *tweet* jika jumlah kata di *tweet* tersebut dibawah 4

Setelah dilakukan pra-proses, data yang bisa digunakan berkurang menjadi sebanyak 779 data tweet.

Berikut merupakan hasil WordCloud per label dari data yang telah di pra-proses:



1. Mempersiapkan data agar bisa dilakukan *training* melakukan *pre-trained model* IndoBERTweet

Pada tahap ini, dilakukan pemrosesan data agar dapat diterima oleh *pre-trained model* IndoBERTweet. Selain itu juga dilakukan *splitting* data *training* dan *testing* dengan besar 20% untuk data *testing*.

1. Mendefinisikan model dan melakukan *training*

Pada tahapan ini, model IndoBERTweet didefinisikan dan dilakukan *training*. *Batch size* yang digunakan adalah 8 dan jumlah *epoch* yang digunakan adalah 25.

1. Evaluasi model

Pada tahap ini, dilakukan evaluasi model dengan beberapa skenario. Berikut merupakan skenario dan hasilnya:

1. Dengan *hashtag* & *username* yang sering muncul, dengan *lemmatizing*, dan dengan *stopwords removal*
   1. *Subset*:

{'eval\_loss': 0.37955841422080994,

'eval\_f1': 0.8061538461538461,

'eval\_roc\_auc': 0.8609000452284035,

'eval\_accuracy': 0.4506172839506173,

'eval\_runtime': 0.7695,

'eval\_samples\_per\_second': 210.533,

'eval\_steps\_per\_second': 27.291,

'epoch': 25.0}

* 1. per label:

precision recall f1-score support

Social 0.85 0.81 0.83 85

Historical 0.77 0.50 0.61 20

Dehumanization 0.73 0.59 0.65 68

Accusation 0.86 0.91 0.88 133

Attack 0.00 0.00 0.00 1

Loyalty 0.00 0.00 0.00 1

Threat 0.94 0.68 0.79 22

micro avg 0.83 0.77 0.80 330

macro avg 0.59 0.50 0.54 330

weighted avg 0.82 0.77 0.79 330

samples avg 0.80 0.77 0.76 330

Accuracy:

Social aspect: 82.7160%

Historical aspect: 91.9753%

Dehumanization: 73.4568%

Accusation in a mirror: 80.2469%

Assertion of Attack Against Women and Girls: 99.3827%

Questioning In-Group Loyalty: 99.3827%

Threat to Group Integrity or Purity: 95.0617%

1. Dengan *hashtag* & *username* yang sering muncul, dengan *lemmatizing*, dan tanpa *stopwords removal*
   1. *Subset*:

{'eval\_loss': 0.4833945333957672,

'eval\_f1': 0.7848932676518883,

'eval\_roc\_auc': 0.8471206996122432,

'eval\_accuracy': 0.4358974358974359,

'eval\_runtime': 0.7208,

'eval\_samples\_per\_second': 216.413,

'eval\_steps\_per\_second': 27.745,

'epoch': 25.0}

* 1. per label:

precision recall f1-score support

Social 0.88 0.72 0.79 89

Historical 0.67 0.29 0.40 21

Dehumanization 0.68 0.64 0.66 56

Accusation 0.82 0.94 0.88 124

Attack 0.00 0.00 0.00 1

Loyalty 0.00 0.00 0.00 1

Threat 0.69 0.50 0.58 18

micro avg 0.80 0.75 0.77 310

macro avg 0.53 0.44 0.47 310

weighted avg 0.79 0.75 0.76 310

samples avg 0.77 0.74 0.73 310

Accuracy:

Social aspect: 78.2051%

Historical aspect: 88.4615%

Dehumanization: 76.2821%

Accusation in a mirror: 79.4872%

Assertion of Attack Against Women and Girls: 99.3590%

Questioning In-Group Loyalty: 99.3590%

Threat to Group Integrity or Purity: 91.6667%

1. Dengan *hashtag* & *username* yang sering muncul, tanpa *lemmatizing*, dan dengan *stopwords removal*
   1. *Subset*:

{'eval\_loss': 0.48915138840675354,

'eval\_f1': 0.7934959349593497,

'eval\_roc\_auc': 0.8559253246753246,

'eval\_accuracy': 0.42948717948717946,

'eval\_runtime': 0.7391,

'eval\_samples\_per\_second': 211.064,

'eval\_steps\_per\_second': 27.06,

'epoch': 25.0}

* 1. per label:

precision recall f1-score support

Social 0.88 0.79 0.83 90

Historical 0.62 0.38 0.47 21

Dehumanization 0.70 0.60 0.65 53

Accusation 0.82 0.90 0.86 125

Attack 0.00 0.00 0.00 1

Loyalty 0.00 0.00 0.00 1

Threat 0.77 0.59 0.67 17

micro avg 0.80 0.76 0.78 308

macro avg 0.54 0.47 0.50 308

weighted avg 0.79 0.76 0.77 308

samples avg 0.77 0.74 0.73 308

Accuracy:

Social aspect: 81.4103%

Historical aspect: 88.4615%

Dehumanization: 77.5641%

Accusation in a mirror: 76.2821%

Assertion of Attack Against Women and Girls: 99.3590%

Questioning In-Group Loyalty: 99.3590%

Threat to Group Integrity or Purity: 93.5897%

1. Dengan *hashtag* & *username* yang sering muncul, tanpa *lemmatizing*, dan tanpa *stopwords removal*
   1. *Subset*:

{'eval\_loss': 0.38720986247062683,

'eval\_f1': 0.8238897396630933,

'eval\_roc\_auc': 0.8773813825235334,

'eval\_accuracy': 0.48148148148148145,

'eval\_runtime': 0.7436,

'eval\_samples\_per\_second': 217.853,

'eval\_steps\_per\_second': 28.24,

'epoch': 25.0}

* 1. per label:

precision recall f1-score support

Social 0.87 0.80 0.83 83

Historical 0.80 0.67 0.73 18

Dehumanization 0.73 0.70 0.72 70

Accusation 0.87 0.95 0.91 131

Attack 0.00 0.00 0.00 1

Loyalty 0.00 0.00 0.00 1

Threat 0.85 0.52 0.65 21

micro avg 0.84 0.81 0.82 325

macro avg 0.59 0.52 0.55 325

weighted avg 0.83 0.81 0.81 325

samples avg 0.83 0.81 0.80 325

Accuracy:

Social aspect: 83.3333%

Historical aspect: 94.4444%

Dehumanization: 75.9259%

Accusation in a mirror: 84.5679%

Assertion of Attack Against Women and Girls: 99.3827%

Questioning In-Group Loyalty: 99.3827%

Threat to Group Integrity or Purity: 92.5926%

1. Tanpa *hashtag* & *username* yang sering muncul, dengan *lemmatizing*, dan dengan *stopwords removal*
   1. *Subset*:

{'eval\_loss': 0.43620964884757996,

'eval\_f1': 0.7640067911714771,

'eval\_roc\_auc': 0.8365893066980024,

'eval\_accuracy': 0.37012987012987014,

'eval\_runtime': 0.7002,

'eval\_samples\_per\_second': 219.951,

'eval\_steps\_per\_second': 28.565,

'epoch': 25.0}

* 1. per label:

Social 0.84 0.75 0.79 76

Historical 1.00 0.32 0.48 19

Dehumanization 0.72 0.71 0.71 58

Accusation 0.78 0.92 0.84 119

Attack 0.00 0.00 0.00 3

Loyalty 0.00 0.00 0.00 1

Threat 0.50 0.10 0.17 20

micro avg 0.78 0.73 0.75 296

macro avg 0.55 0.40 0.43 296

weighted avg 0.77 0.73 0.72 296

samples avg 0.75 0.74 0.72 296

Accuracy:

Social aspect: 80.5195%

Historical aspect: 91.5584%

Dehumanization: 78.5714%

Accusation in a mirror: 74.0260%

Assertion of Attack Against Women and Girls: 98.0519%

Questioning In-Group Loyalty: 99.3506%

Threat to Group Integrity or Purity: 87.0130%

1. Tanpa *hashtag* & *username* yang sering muncul, dengan *lemmatizing*, dan tanpa *stopwords removal*
   1. *Subset*:

{'eval\_loss': 0.30863404273986816,

'eval\_f1': 0.7785016286644951,

'eval\_roc\_auc': 0.831002423093131,

'eval\_accuracy': 0.4472049689440994,

'eval\_runtime': 0.7574,

'eval\_samples\_per\_second': 212.577,

'eval\_steps\_per\_second': 27.727,

'epoch': 25.0}

* 1. per label:

Social 0.87 0.72 0.79 94

Historical 1.00 0.19 0.32 31

Dehumanization 0.75 0.38 0.51 55

Accusation 0.86 0.88 0.87 129

Attack 0.00 0.00 0.00 0

Loyalty 0.00 0.00 0.00 3

Threat 1.00 0.12 0.22 24

micro avg 0.86 0.63 0.73 336

macro avg 0.64 0.33 0.39 336

weighted avg 0.86 0.63 0.69 336

samples avg 0.79 0.65 0.69 336

Accuracy:

Social aspect: 77.6398%

Historical aspect: 84.4720%

Dehumanization: 74.5342%

Accusation in a mirror: 79.5031%

Assertion of Attack Against Women and Girls: 100.0000%

Questioning In-Group Loyalty: 98.1366%

Threat to Group Integrity or Purity: 86.9565%

1. Tanpa *hashtag* & *username* yang sering muncul, tanpa *lemmatizing*, dan dengan *stopwords removal*
   1. *Subset*:

{'eval\_loss': 0.5072057843208313,

'eval\_f1': 0.7878787878787878,

'eval\_roc\_auc': 0.8539289527241184,

'eval\_accuracy': 0.4129032258064516,

'eval\_runtime': 0.7945,

'eval\_samples\_per\_second': 195.089,

'eval\_steps\_per\_second': 25.173,

'epoch': 25.0}

* 1. per label:

Social 0.86 0.79 0.82 82

Historical 0.68 0.72 0.70 18

Dehumanization 0.74 0.67 0.70 63

Accusation 0.81 0.95 0.88 120

Attack 0.00 0.00 0.00 2

Loyalty 0.00 0.00 0.00 2

Threat 0.47 0.41 0.44 22

micro avg 0.78 0.79 0.78 309

macro avg 0.51 0.51 0.51 309

weighted avg 0.77 0.79 0.77 309

samples avg 0.78 0.80 0.76 309

Accuracy:

Social aspect: 81.9355%

Historical aspect: 92.9032%

Dehumanization: 76.7742%

Accusation in a mirror: 79.3548%

Assertion of Attack Against Women and Girls: 98.7097%

Questioning In-Group Loyalty: 98.7097%

Threat to Group Integrity or Purity: 85.1613%

1. Tanpa *hashtag* & *username* yang sering muncul, tanpa *lemmatizing*, dan tanpa *stopwords removal*
   1. *Subset*:

{'eval\_loss': 0.4903532564640045,

'eval\_f1': 0.7901234567901235,

'eval\_roc\_auc': 0.8553903345724907,

'eval\_accuracy': 0.42857142857142855,

'eval\_runtime': 0.7847,

'eval\_samples\_per\_second': 205.183,

'eval\_steps\_per\_second': 26.763,

'epoch': 25.0}

* 1. per label:

precision recall f1-score support

Social 0.81 0.80 0.81 91

Historical 0.70 0.50 0.58 28

Dehumanization 0.70 0.72 0.71 53

Accusation 0.81 0.96 0.88 123

Attack 0.00 0.00 0.00 0

Loyalty 0.00 0.00 0.00 3

Threat 0.77 0.45 0.57 22

micro avg 0.78 0.79 0.79 320

macro avg 0.54 0.49 0.51 320

weighted avg 0.77 0.79 0.77 320

samples avg 0.77 0.80 0.76 320

Accuracy:

Social aspect: 78.2609%

Historical aspect: 87.5776%

Dehumanization: 80.7453%

Accusation in a mirror: 79.5031%

Assertion of Attack Against Women and Girls: 100.0000%

Questioning In-Group Loyalty: 98.1366%

Threat to Group Integrity or Purity: 90.6832%

1. *Model summary*

Tahapan ini menunjukkan arsitektur model yang digunakan.

1. Inferensi model

Pada tahapan ini, dilakukan *testing* menggunakan kalimat apapun yang ingin di-*predict*. Selain itu, data testing beserta label hasil prediksinya disimpan dalam bentuk .csv.

**Kesimpulan**

Dari hasil data diatas, *subset accuracy* terbaik didapatkan pada skenario 4, yaitu di akurasi 48,148%, yaitu dengan menggunakan *hashtag* & *username* yang sering muncul, tanpa *lemmatizing*, dan tanpa *stopwords removal*, sedangkan untuk *subset accuracy* terburuk didapatkan pada skenario 4, yaitu di akurasi 37,013%, yaitu tanpa menggunakan *hashtag* & *username* yang sering muncul, dengan *lemmatizing*, dan dengan *stopwords removal*