LAPORAN TUGAS TEXT CLASSIFICATION

Disusun Oleh :

Gery Nugroho (1301170116)

1. **Mengubah Format Penyimpanan menjadi CSV**

karena kondisi datanya dalam bentuk pickle agar dapat dilihat dan dibaca dengan mudah melalui Ms. Excel. maka data dapat diubah ke dalam bentuk CSV.

```
## simpan dalam bentuk csv

df.to_csv('Data/df.csv',index=False)
np.savetxt('Data/features_train.csv', features_train, delimiter=",")
np.savetxt('Data/labels_train.csv', labels_train, delimiter=",")
np.savetxt('Data/features_test.csv',features_test,delimiter=",")
np.savetxt('Data/labels_test.csv',labels_test,delimiter=",")
```

untuk data df.picke dapat diubah menggunakan library pandas nya langsung.

sedangkan untuk yang lainnya karena tidak dalam bentuk format DataFrame maka dapat menggunakan library np.savetxt untuk menyimpan datanya

2. **Mengubah Features**

dengan menggunakan algoritma Logistic Regression. didapatkan hasil berikut.

	Model	Training Set Accuracy	Test Set Accuracy
0	Logistic Regression Normal	0.987837	0.943114
1	Logistic Regression Tanpa Lematisasi	0.985722	0.946108
2	Logistic Regression Tanpa Normalisasi	0.986779	0.934132
3	Logistic Regression Tanpa Stop Word	0.976732	0.913174

- dengan melakukan proses lematisasi, normalisasi, dan stop word. didapatkan hasil akurasi sebesar **94.31%** terhadap pengujian data test dan **98.78%** terhadap data training.
- dengan melakukan proses tanpa lematisasi maka didapatkan hasil akurasi sebesar **94.61%** terhadap pengujian data test **lebih besar 0.30%** terhadap proses normal sedangkan terhadap pengujian data training didapatkan **98.57%** terdapat **penurunan** sebesar **0.21%.** dari sini dapat kita ketahui jika tanpa proses lematisasi maka untuk akurasinya akan **menurun** apabila menggunakan data yang sudah ada sedangkan akurasinya aakn **meningkat** apabila menggunakan data yang tidak ada dalam model
- dengan melakukan proses tanpa normalisasi maka didapatkan hasil akurasi sebesar **93.41%** terhadap pengujian data test **lebih kecil 1.10%** terhadap proses normal sedangkan terhadap pengujian data training didapatkan **98.67%** terdapat **penurunan** sebesar **0.11%.** dari sini dapat kita ketahui jika tanpa proses normalisasi maka untuk

akurasinya akan **menurun** dibanding menggunakan proses normal baik itu menggunakan data yang sudah ada dan data yang belum ada

- dengan melakukan proses tanpa stop word maka didapatkan hasil akurasi sebesar **91.31%** terhadap pengujian data test **lebih kecil 3.0%** terhadap proses normal sedangkan terhadap pengujian data training didapatkan **97.67%** terdapat **penurunan** sebesar **1.11%.** dari sini dapat kita ketahui bahwa tanpa stop word maka hasil akurasi akan lebih kecil daripada proses normal.

3. Membedakan Max Features

dengan menggunakan algoritma logistic regression maka didapatkan hasil berikut terhadap perbedaan max features.

	Model	Training Set Accuracy	Test Set Accuracy
0	Max Features 300	0.987837	0.943114
1	Max Features 200	0.961396	0.895210
2	Max Features 400	0.984664	0.925150

dari hasil tersebut didapatkan bahwa hasil akurasi tidak bergantung pada banyak atau sedikitnya max features yang dimiliki oleh model.

4. Membandingkan Beberapa Algoritma

algoritma yang akan dibandingkan adalah algoritma Logistic Regresssion dan Random Forest. berikut adalah hasilnya.

	Model	Training Set Accuracy	Test Set Accuracy	Random Search Accuracy	Grid Search Accuracy
0	Logistic Regression	0.987837	0.943114	0.958752	0.970133
0	Random Forest	1.000000	0.934132	0.942359	0.941867

dari hasil tersebut diliat bahwa Random Forest benar semua sehingga bisa jadi terjadi overfitting. namun dalam hal lainnya algoritma logistic regression memiliki akurasi yang lebih tinggi untuk beberapa test akurasi. dan juga terdapat kekurangan dari random forest yaitu dengan lebih lama dalam proses trainingnya.

5. Menggunakan Teks Bahasa Indonesia

bagian yang perlu penyesuaian adalah bagian feature engineering nya yaitu Normalisasi, Lematisasi, dan Stop word.

pada bagian normalisasi salah satu contohnya adalah menghilangkan (‘s) untuk kepunyaan karena pada kata bahasa indonesia tidak terdapat (‘s).

pada bagian lematisasi yang mengubah kata dasar ke kata bahasa indonesianya.

stopword juga perlu disesuaikan sesuai bahasa indonesia.