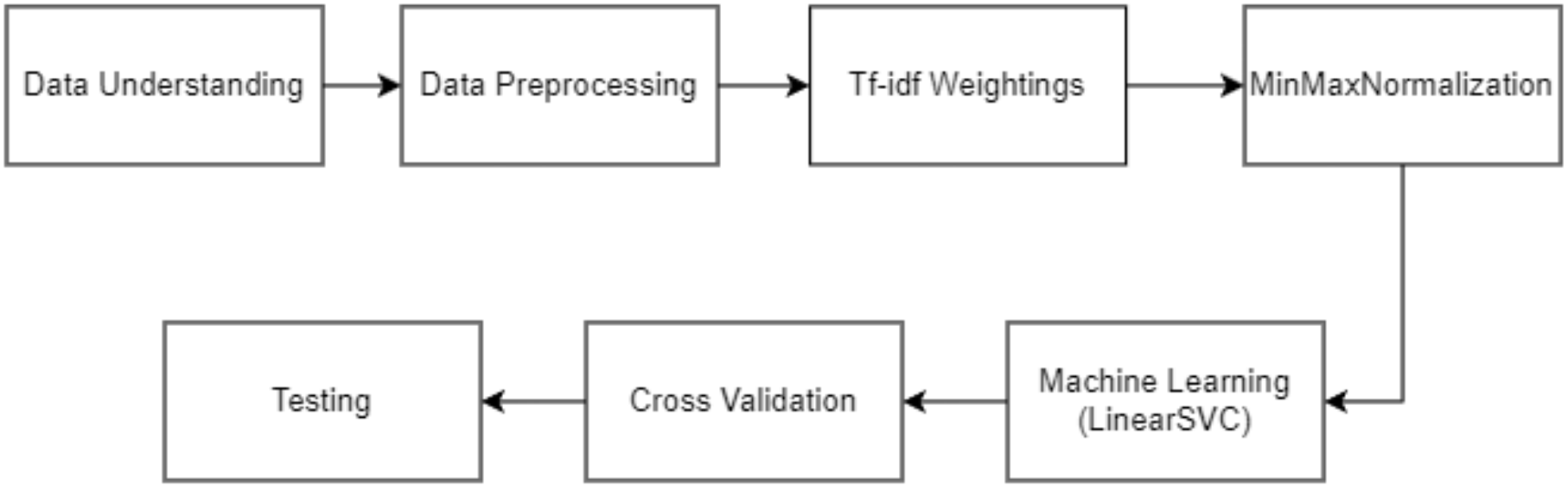
Emotion Recognition by Text

Rizky Alif Ramadhan
(19/446785/TK/49890)

Background and Purpose

Pemakaian media sosial terkadang menimbulkan dampak negatif bagi kesehatan mental. Kegiatan *chatting* untuk berbagi cerita dapat menjadi solusi atas masalah kesehatan mental yang dialami. Sebagai orang yang mendengarkan, terkadang kita salah mengartikan emosi dari lawan bicara kita. Dengan melakukan analisis sentimenmenggunakan model machine learning ini mesin dapat mengetahui emosi dari lawan bicara kita dengan baik, mesin dapat memberi tahu kita, sehingga kita dapat mengambil tindakan yang tepat.

Sentiment Analysis Pipeline



Data Understanding

```
df
```

	sentence	label
0	i didnt feel humiliated	sadness
1	i can go from feeling so hopeless to so damned...	sadness
2	im grabbing a minute to post i feel greedy wrong	anger
3	i am ever feeling nostalgic about the fireplac...	love
4	i am feeling grouchy	anger
...	...	...
17995	im having ssa examination tomorrow in the morn...	sadness
17996	i constantly worry about their fight against n...	joy
17997	i feel its important to share this info for th...	joy
17998	i truly feel that if you are passionate enough...	joy
17999	i feel like i just wanna buy any cute make up ...	joy

18000 rows × 2 columns

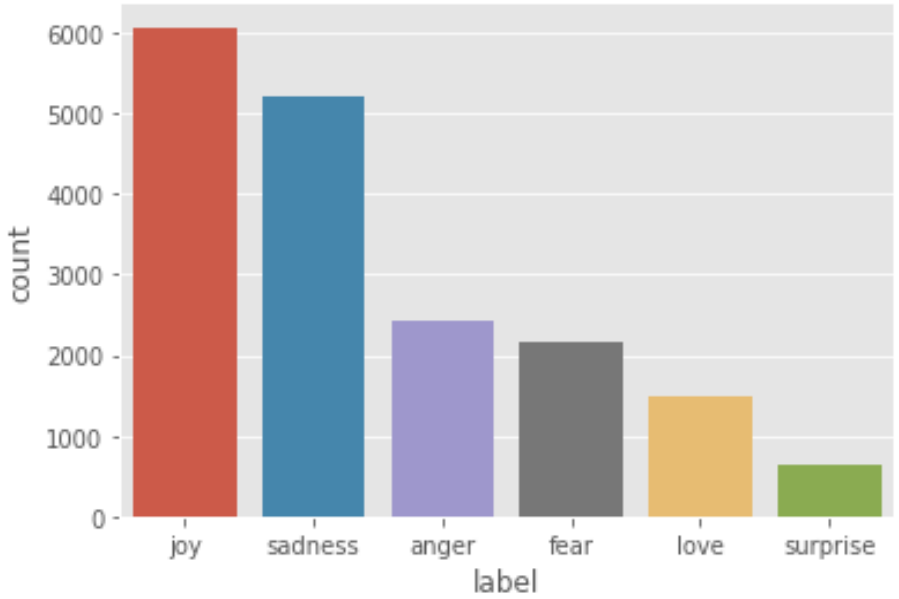
```
df_test
```

	sentence	label
0	im feeling rather rotten so im not very ambiti...	sadness
1	im updating my blog because i feel shitty	sadness
2	i never make her separate from me because i do...	sadness
3	i left with my bouquet of red and yellow tulip...	joy
4	i was feeling a little vain when i did this one	sadness
...	...	...
1995	i just keep feeling like someone is being unki...	anger
1996	im feeling a little cranky negative after this...	anger
1997	i feel that i am useful to my people and that ...	joy
1998	im feeling more comfortable with derby i feel ...	joy
1999	i feel all weird when i have to meet w people ...	fear

2000 rows × 3 columns

```
missing_percentage(df)
```

	Total	Percent
label	0	0.0
sentence	0	0.0



Imbalanced Class

Data Preprocessing

- Label Encoding : Melakukan Encoding pada Data Kategorik menjadi "Numerik"

	label	label_enc
0	sadness	4
2	anger	0
3	love	3
6	surprise	5
7	fear	1
8	joy	2

Tf-idf Weighting + Removing Stopwords

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

Tf-idf Weighting

- Misalnya kita ingin menghitung tfidf kata wrong pada dokumen kedua yaitu *"im grabbing a minute to post i feel greedy wrong"*

- Terdapat 1 kata "wrong" dari 6 kata pada dokumen tersebut (stop words tidak dihitung). Maka dari itu tf = 1/6 = 0.1667

- Terdapat 78 dokumen yang mengandung kata "wrong" dari 18000 dokumen. Maka idfnya adalah Log(18000/78) = 2.363

- Nilai dari pembobotan tfidnya adalah 0.1667 x 2.363 = 0.393833

MinMaxScaler for Normalization

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Digunakan agar tidak terjadi kejomplangan nilai antara kolom satu dengan kolom lainnya

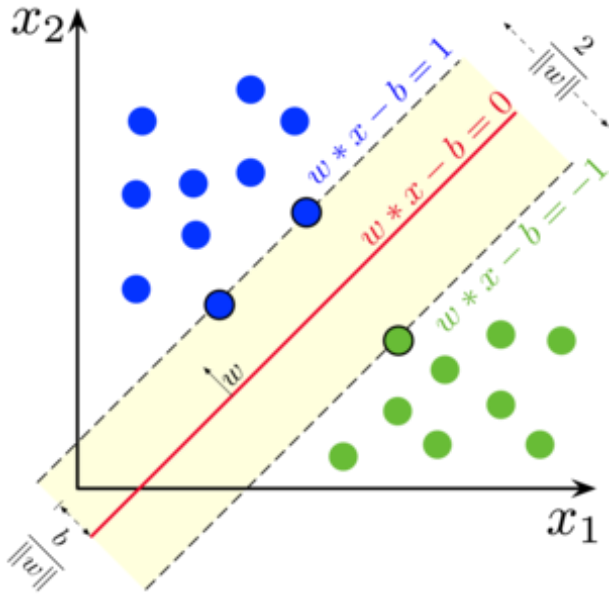
MinMaxScaler for Normalization

Misalnya kata "wrong" pada dokumen kedua, sebelumnya memiliki nilai tfid = 0.393833. Kolom "wrong" memiliki nilai minimal 0 dan nilai maksimal 0.8. Maka dari itu setelah di normalisasi kolom "wrong" pada dokumen kedua akan bernilai :

(0.393833 - 0) / (0.8 - 0) = 0.49229125

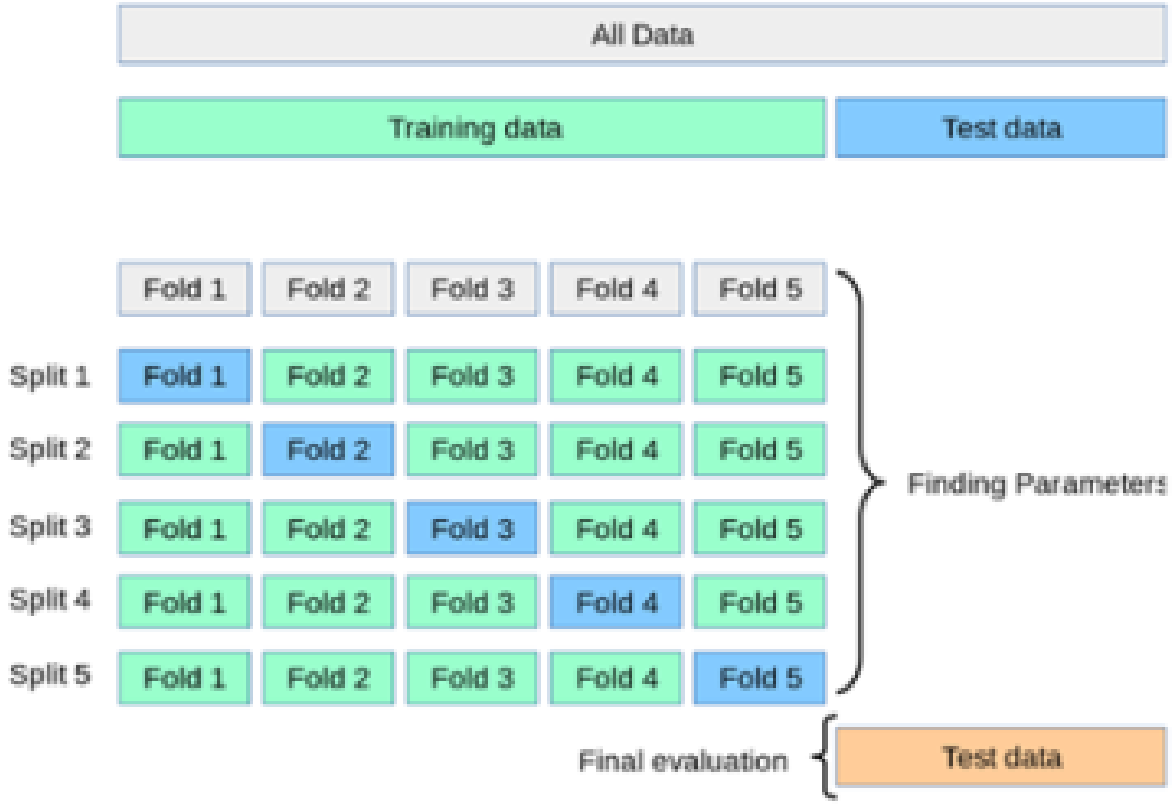
LinearSVC

Support Vector Machine (SVM) adalah pengklasifikasi diskriminatif yang secara formal didefinisikan oleh hyperplane pemisah. Dengan kata lain, diberikan data pelatihan berlabel (pembelajaran yang diawasi), algoritme menghasilkan hyperplane optimal yang mengkategorikan contoh-contoh baru. Dalam ruang dua dimensi hyperplane ini adalah sebuah garis yang membagi sebuah bidang menjadi dua bagian dimana pada setiap kelasnya terletak pada kedua sisinya. Pada kasus ini kita akan menggunakan **hyperplane linear.**



5-Fold Cross Validation

Membuat skenario validasi sebagimana gambar dibawah :



Didapat hasil rata-rata F1 Score :

model_name	Mean F1_Macro	Standard deviation
LinearSVC	0.841194	0.00941

Classification Report

Setelah melakukan validasi, kita akan mencobanya pada data tes, didapat :

	precision	recall	f1-score	support
anger	0.86	0.86	0.86	275
fear	0.86	0.83	0.85	224
joy	0.89	0.91	0.90	695
love	0.73	0.74	0.74	159
sadness	0.92	0.91	0.91	581
surprise	0.71	0.68	0.70	66
accuracy			0.87	2000
macro avg	0.83	0.82	0.83	2000
weighted avg	0.87	0.87	0.87	2000

Dengan :

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

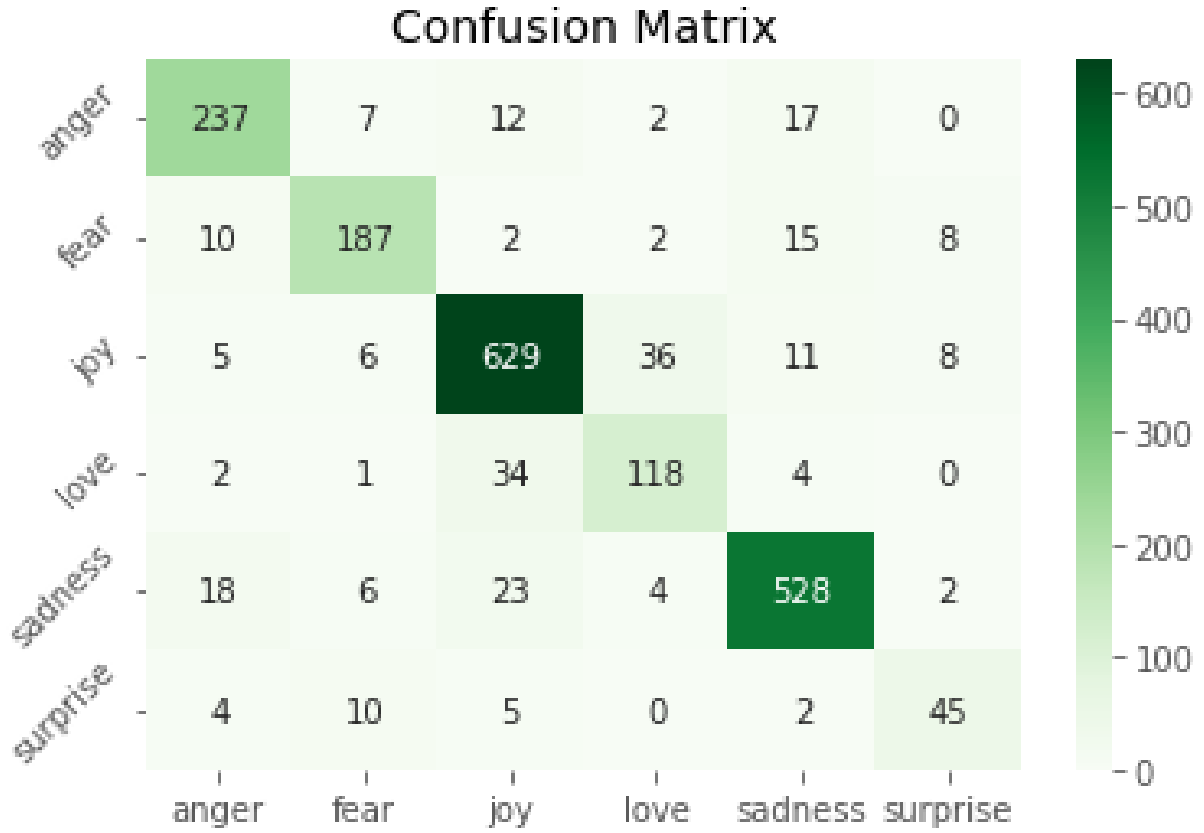
$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}$$

$$\text{F}_1 \text{ score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

tp : true positif, fp : false positif
Tn : true negatif, fn = false negatif

Confusion Matrix



Conslusion

- Telah berhasil dibuat model LinearSVC untuk melakukan Emotion Recognition dari Text dengan nilai rata-rata F1 Score = 0.841194. Pada data tes didapat F1 Score dengan bobot = 0.87
- Data yang dipakai imbalance pada kelas-kelasnya, ini akan menimbulkan bias pada kelas-kelas yang memiliki jumlah sedikit. Dapat dilihat pada kelas "joy" dan "sadness" yang merupakan dua kelas terbanyak, F1 Scorenya sangat baik yaitu > 0.90. Sementara kelas yang lainnya kurang baik.

Conslusion

- Pada confusion matrix dapat dilihat bahwa banyak emosi "joy" yang diprediksi "love", begitu juga sebaliknya. Ini mengindikasikan bahwa kedua emosi tersebut mempunyai hubungan yang dekat.
- Emosi "anger" dan "sadness" juga demikian, namun tidak sedekat emosi "joy" dan "love"

Recommedation

- Kita dapat menambah data-data pada kelas yang imbalance agar tidak terjadi bias pada model yang dibuat.
- Melakukan multi-class classification merupakan kegiatan yang menantang, untuk memperbaiki akurasi kita dapat menjadikannya binary class saja, emosi positif dan negatif. (Love, joy, surprise : Positif), (Sadness, fear, anger : Negatif).
- Untuk memperbaiki model, dapat dilakukan hyperparameter tuning untuk mencari parameter terbaik dari LinearSVC yang digunakan untuk memodelkan kasus ini.
- Preprosesing seperti Lemmatization dan Stemming, mungkin akan memperbaiki akurasi, mungkin juga tidak.