TUGAS

PEMROSESAN BAHASA ALAMI

SENTIMENT ANALYSIS SUPERVISED LEARNING MENGGUNAKAN ALGORITMA RANDOM FOREST TERHADAP UNGGAHAN TWITTER #PERCUMALAPORPOLISI

OLEH :

ANNIDA NUR ISAMI

19102240

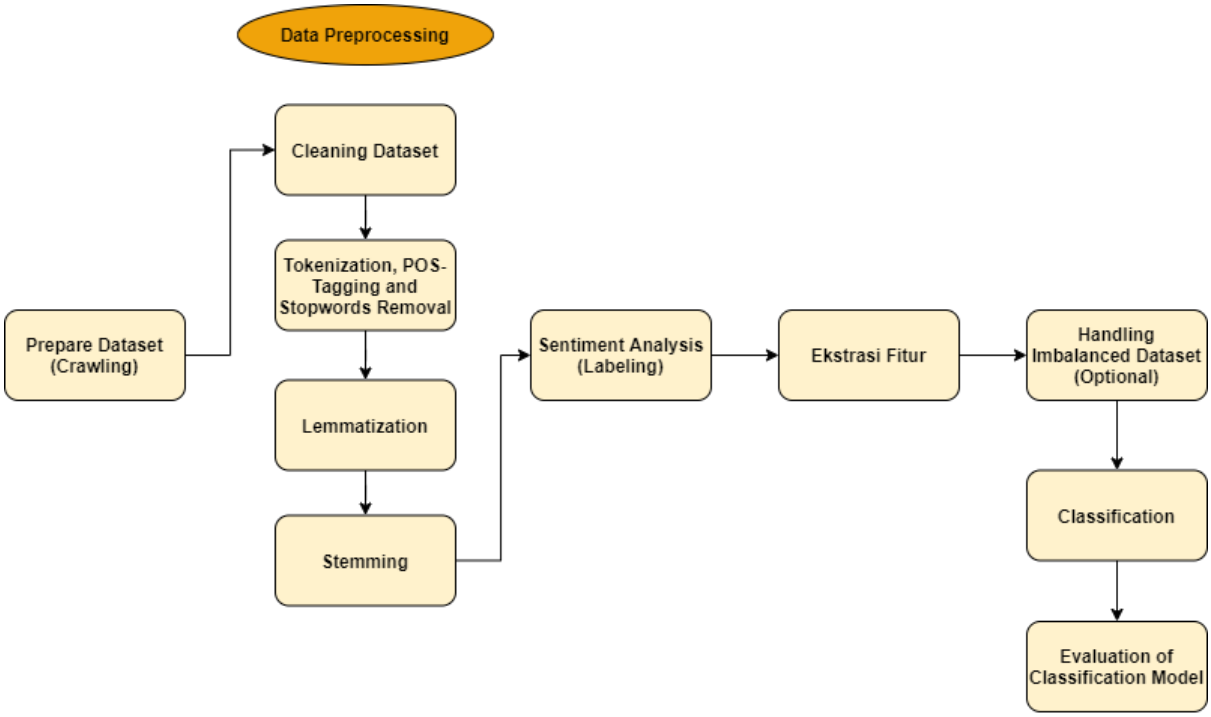
INSTITUT TEKNOLOGI TELKOM PURWOKERTO

FAKULTAS INFORMATIKA

PROGRAM STUDI S1 TEKNIK INFORMATIKA

2021

I. ALUR



Gambar 1. Alur Sentiment Analyis Supervised Learning

Setelah memiliki file csv dataset berisi unggahan twitter dengan #PercumaLaporPolisi yang didapatkan dengan crawling data menggunakan twint *library*, dapat langsung melakukan alur *sentiment analysis supervised learning* yaitu :

- ***Data Preprocessing***

  Tahap *data preprocessing* dilakukan untuk memastikan bahwa data yang akan diproses selanjutnya bersih. Bersih yang dimaksud di sini adalah data yang akan diproses memiliki format yang benar, karena seringkali data yang didapatkan tidak lengkap, tingkah konsisten, dan cenderung memiliki beberapa kesalahan. Langkah dalam *data preprocessing* antara lain :

  - ***Cleaning Dataset*** : menghilangkan karakter non-alfabet
  - ***Tokenization*** : memecah dokumen menjadi bagian atau token
  - ***POS-Tagging*** : menentukan atribut/tipe dari sebuah kata
  - ***Stopword Removal*** : menghilangkan kata tidak penting/tidak mengubah arti
  - ***Lemmatization*** : mengubah kata ke kata baku dengan konteks kalimat
  - ***Stemming(Sastrawi)*** : mengubah kata ke kata baku tanpa konteks kalimat.

- ***Sentiment Analysis / Labeling***

  Setelah dataset bersih, maka tahap *sentiment analysis* atau *labeling* dapat dilakukan. Di tahap ini seluruh unggahan twitter dengan #PercumaLaporPolisi akan dikategorikan ke dalam unggahan positif, negatif atau netral. Metode *sentiment analysis*/labeling yang dapat digunakan ada beberapa seperti TextBlob, Vader, dan SentiwordNet. Metode-metode tersebut memiliki kelebihannya masing-masing, sedangkan untuk tugas kali ini menggunakan TextBlob.

- **Ekstrasi Fitur**

  Merupakan tahap yang sangat penting dalam pengenalan pola klasifikasi nantinya. Secara umum tahap ini bertujuan untuk memperoleh informasi yang terkandung dalam suatu citra untuk kemudian dijadikan sebagai acuan untuk membedakan antara citra yang satu dengan citra yang lain (Pamungkas Adi, 2017). Namun dalam kasus ini, ekstrasi fitur digunakan untuk mencari bobot dari sebuah kata (seberapa sering muncul). Teknik ekstrasi fitur yang dapat digunakan adalah Bag-of-Words atau TF-IDF. Pada tugas kali ini ekstrasi fitur yang akan digunakan adalah TF-IDF.

- ***Handling Imbalanced Dataset***

  Untuk tahap ini tidak bersifat wajib atau *optional*, perlu dilakukan hanya saat hasil dari *sentiment analysis*/labeling tidak seimbang dimana kuantitas dari salah satu kategori jauh lebih tinggi atau mendominasi dari kategori yang lain. Dampak yang ditimbulkan dari penggunaan data yang tidak seimbang memang bersifat implisit (tidak langsung saat model klasifikasi dibangun dan dijalankan), tetapi klasifikasi yang dihasilkan tidak benar. Karena jika sebuah model klasifikasi dibangun menggunakan data yang tidak seimbang, maka hasil dari model klasifikasi tersebut cenderung menghasilkan kategori yang mayoritas dan berdampak pada tingkat akurasi yang akan selalu tinggi. Terdapat 3 pendekatan untuk menangani data yang tidak seimbang yaitu level-data (random-oversampling, random-undersampling, dan FSMOTE), level-algoritmik, serta gabungan. Namun pada tugas ini, handling imbalanced dataset dengan pendekatan level-data algoritma FSMOTE.

- ***Clasiffication***

  Setelah dipastikan data yang akan diklasifikasikan seimbang, maka proses klasifikasi dapat langsung dilakukan. Banyak algoritma yang dapat digunakan untuk klasifikasi seperti Naive Bayes, SVM, Random Forest, TF, Neutral Network, dan lain-lain. Di tugas kali ini algoritma yang digunakan adalah Random Forest dengan perbandingan data *train* dan *test*-nya adalah 7:3 atau 70% : 30%.

- ***Evaluation of Classification Model***

  Tahap ini perlu dilakukan untuk mendapatkan informasi tentang kinerja dari model klasifikasi yang telah dibangun dan dijalankan. Ada beberapa istilah yang akan ditampilkan di evaluasi model klasifikasi yaitu :

  - ***Accuracy***

    Menggambarkan seberapa akurat model yang telah dibuat untuk dapat mengklasifikasikan data dengan benar, dengan kata lain tingkat kedekatan nilai prediksi dengan nilai sebenarnya. Merupakan rasio prediksi benar positif (*true positive*) dengan keseluruhan data, dapat dirumuskan :

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

  - ***Precision***

    Menggambarkan tingkat keakuratan antara data prediksi benar positif (*true positive*) yang diminta dengan hasil prediksi yang diberikan model klasifikasi. Merupakan rasio benar positif (*true positive*) dengan keseluruhan hasil yang diprediksi positif. Dapat dirumuskan sebagai berikut :

$$precision = \frac{TP}{TP + FP}$$

  - ***Recall***

    Sering juga disebut *sensitivity (true positive rate).* Menggambarkan keberhasilan model dalam menentukan kembali sebuah informasi. Merupakan

rasio prediksi benar positif (*true positive*) dengan keseluruhan data yang benar positif. Dapat dirumuskan sebagai berikut :

$$recall = \frac{TP}{TP + FN}$$

- ***F1-Score***

  Perbandingan rata-rata *presicion* dan *recall*. Semakin dekat nilai *f1-score* yang didapatkan dengan 1, maka semakin baik kinerja model klasifikasi.

- ***Support***

  Jumlah aktual dataset, dan diagnosis proses evaluasi kinerja model klasifikasi.

II. HASIL

Setelah tahap *data preprocessing*, hasil *sentiment analysis* menggunakan TextBlob pada dataset unggahan twitter dengan #PercumaLaporPolisi menunjukan bahwa :

- 167 tweet bersifat positif
- 103 tweet bersifat negatif
- 4741 tweet bersifat netral

Dari *wordcloud* setiap kategori, terlihat bahwa kata yang paling mendominasi di setiap kategori adalah ‘percumalaporpolisi’. Seperti di bawah ini :

- Kategori Positif

- Kategori Negatif

- Kategori Netral

Sedangkan pada tahap ekstrasi fitur menggunakan TF-IDF didapatkan hasil pembobotan untuk setiap kata yang ada dalam dataset, kata dengan bobot paling :

- Tinggi : ‘zul’ muncul sebanyak 7810 kali
- Rendah : ‘aaaaaaaaah’ muncul sebanyak 0 kali

Karena hasil *sentiment analysis* tidak seimbang antara jumlah tweet setiap kategori, maka dilakukan *handling imbalanced dataset* dengan SMOTE dengan hasil jumlah tweet masing-masing kategori adalah 4741.

Setelah dataset seimbang dan dilakukan klasifikasi dataset dengan algoritma random forest didapatkan *classification report, f1-score* dan akurasi dari model klasifikasi yang telah dibuat dan dijalankan. Sebagai berikut :

```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1413
           1       0.99      1.00      1.00      1438
           2       1.00      1.00      1.00      1416

    accuracy                           1.00      4267
   macro avg       1.00      1.00      1.00      4267
weighted avg       1.00      1.00      1.00      4267

F1-Score :  [0.99929229 0.99653019 0.99716914]
Accuracy :  0.9976564330911648
```

- Akurasi yang didapatkan sekitar 0.99. Angka tersebut sangat dekat dengan 1 yang merupakan indeks terbesar akurasi, sehingga dengan akurasi 0.99 dapat dikatakan model klasifikasi yang dibuat dan dijalankan sudah cukup baik
- F1-Score yang didapat semua berkisar 0.99 yang merepresentasikan rata-rata precision dan recall mendekati 1. Dengan angka tersebut dapat disimpulkan bahwa model klasifikasi yang dibuat memiliki kinerja yang cukup baik.