Laporan NLP Klasifikasi Teks Pada Forum Kaskus

Nama : Fikri Rozan Imadudin/1301150768,Riko Bintang Purnomo Putra/1301154714, Jeqwalin Claudya Patandianan/1301150737
Kelas : ICM-39-GAB

Analisis dan strategi penyelesaian masalah.

Tujuan dan Masalah :

1. Diberikan data komentar kaskus gender.csv dan komentar kaskus gender test.csv yang memiliki label yaitu Komentar dan Gender. komentar kaskus gender akan digunakan sebagai data training dan komentar kaskus gender test akan digunakan sebagai data testing.
2. Pada tugas klasifikasi teks kali ini akan dibuat sebuah sistem klasifikasi menggunakan metode metode *Bagging*, salah satu teknik *Ensemble Learning*, berbasis *Multinomial Naïve Bayes* untuk menentukan kelas/label data uji dalam kaskus gender test.

Teori dan Kajian :

Data

Data diambil dari forum Kaskus.com pada forum *ask da boys* dan *ask da girls* secara manual dengan pelabelan Pria atau Wanita. Data yang diambil memiliki jumlah data 105 untuk data training dengan 55 Wanita dan 50 Pria dan 15 untuk data testing 8 Pria dan 7 Wanita.

Gender Klasifikasi

Gender klasifikasi adalah secara otomatis dengan metode *machine learning* membedakan suatu gender **Pria** atau **Wanita** berdasarkan isi dari sebuah data corpus yang telah dilabelkan.

Preprocessing

- *Stopwords* merupakan kata umum yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna. Contoh stopwords dalam bahasa Indonesia meliputi : yang, pada, namun, menurut, antara, dulu, ia.
- *Casefolding* merupakan teknik merubah huruf besar menjadi huruf kecil di dalam corpus.
- *NumberRemoval* merupakan teknik menghapus nomor dalam corpus

MNNB dan *Bagging*

Naïve Bayes Classifier merupakan sebuah pengklasifikasi probabilitas sederhana yang mengaplikasikan *Teorema Bayes* dengan asumsi ketidaktergantungan yang tinggi. *Teorema Bayes* adalah teorema yang dipakai dalam statistika untuk menghitung peluang untuk suatu hipotesis. Umumnya terdapat tiga model distribusi yaitu *Bernoulli*, *Multinomial,* dan *Poisson* ketiga model ini digunakan sebagai *classifier* yaitu *Bernouli Naïve Bayes*, *Multinomial Naïve Bayes* dan *Poisson Naïve Bayes*. *Multinomial Naïve Bayes* memperhitungkan jumlah kata dalam dokumen sehingga mengasumsikan independensi kemunculan kata dalam dokumen.Dengan asumsi ini menunjukan bahwa kemungkinan tiap kejadian kata dalam dokumen adalah bebas tidak memperhitungkan urutan kata dan konteks kata dalam dokumen.

Dengan persamaan:

$$P(w_i,c_k) = \frac{1 + \sum_{j=1}^{N} N_{ij}\delta_{jk}}{|V| + \sum_{i=1}^{|V|}\sum_{n=1}^{N} N_{in}\delta_{jk}} \quad (1)$$

Dengan:

N adalah jumlah dari dokumen.
δ_{jk} bernilai 1 jika dokumen ke j milik kelas c_k dan bernilai 0 jika sebaliknya.
N_{ij} adalah jumlah fitur w_i yang terjadi pada dokumen ke j+1 untuk menghindari nilai nol atau adanya satu kemungkinan.

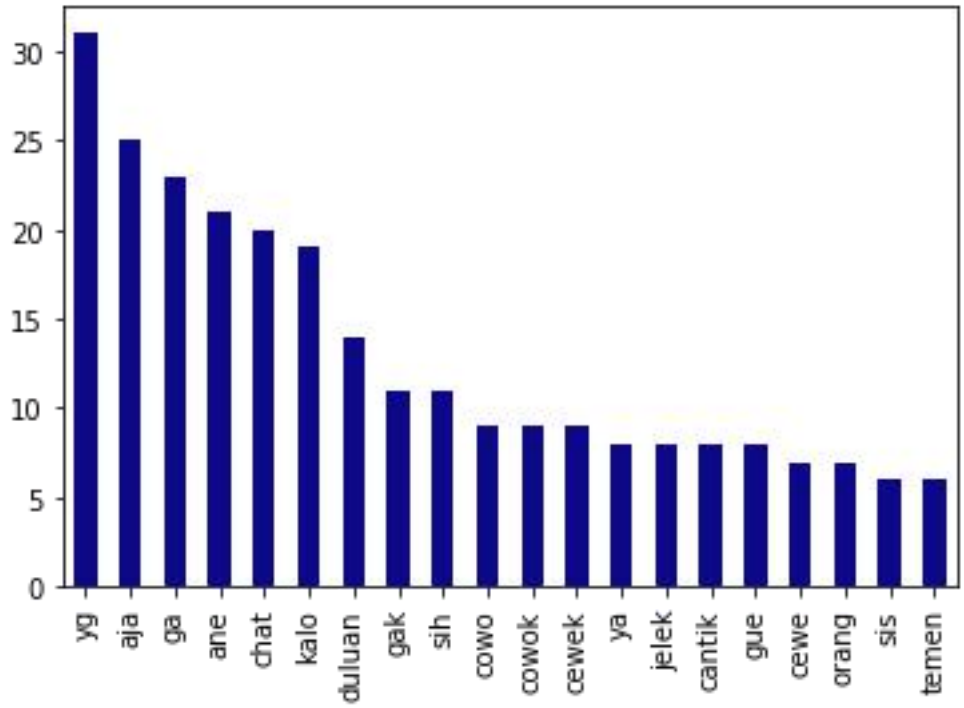
Bagging merupakan salah satu dari *Ensemble Learning* yang digunakan untuk mengurangi variansi, meningkatkan akurasi algoritma klasifikasi dan dapat menghindari *overfitting*. Idenya yaitu membagi data kedalam beberapa *subspace* lalu mengkombinasikan data secara acak dengan ukuran yang sama dengan data asli dalam masing-masing subspace sehingga didapatkan dari masing-masing *subspace* rata-rata untuk prediksi atau klasifikasi yang akan digunakan nanti.

Metode Penelitian :
Metodologi yang digunakan terdiri dari beberapa tahap yaitu memasukan dataset, *Preprocessing*, memisahkan dataset menjadi dua bagian yaitu (*x_train, y_train, x_test, y_test*), perancangan sistem, pengujian model dan analisa hasil. Akan digunakan sebuah library sklearn BaggingClassifier yang merupakan metode bagging dan MultinomialNB sebagai naive bayes.

Analisis dan Hasil :
```
Data yang digunakan mengandung 105 data yaitu :
wanita    55
pria      50
```

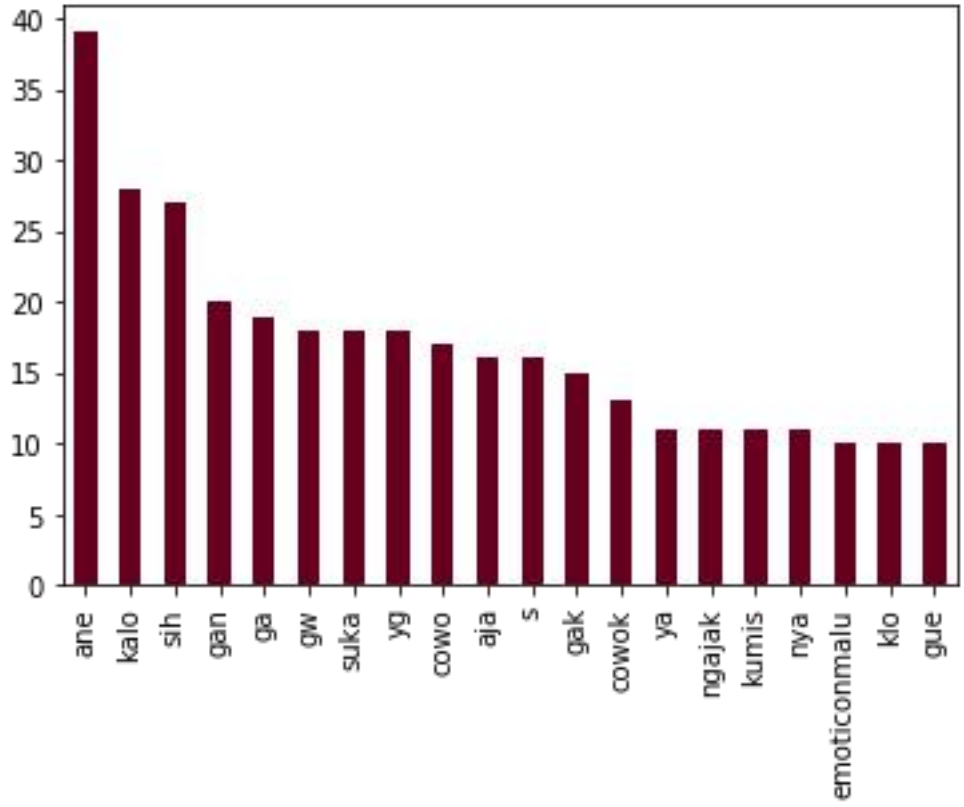
Kita akan melihat kata apa yang sering muncul di dalam corpus berdasarkan Gender pria atau wanita.

1 PlotBar Pria



Kata yang paling sering muncul dalam komentar pria yaitu yg, aja, ga, ane

2 PlotBar Wanita



Kata yang paling sering muncul dalam komentar pria yaitu ane,kalo,sih.

Experiment dilakukan terhadap *MNNB* tanpa menggunakan *Bagging* dan *MNNB* dengan *Bagging.* Pada tabel berikut berupa hasil yang telah diterapkan sebagai berikut:

Algoritma	Akurasi
Naïve Bayes tanpa menggunakan *Bagging*	85. 71 %
Naïve Bayes dengan *Bagging*	92. 86 %

Hasil experiment yang ada pada tabel berikut menyatakan bahwa penerapan teknik bagging pada *Naïve Bayes* meningkatkan nilai akurasi sebesar 7.51 %.

Kesimpulan :

Pada percobaan kali ini bahwa berdasarkan kata yang sering muncul umumnya bahwa komentar di dalam Kaskus banyak mengandung kata yang disingkat-singkat dan bahasa yang tidak baku seperti yang menjadi yg, kalau menjadi klo, gw, cowok, cewek dll. Didapatkan kesimpulan bahwa pengklasifikasian sebuah gender pada komentar mendapatkan hasil yang cukup baik walaupun hanya dengan 105 corpus dengan maksimum akurasi 92.86 %.

Referensi

https://www.kaskus.co.id/forum/105/ask-da-girls/?ref=postlist&med=breadcrumb
https://www.kaskus.co.id/forum/114/ask-da-boys/?ref=postlist&med=breadcrumb
https://scikit-learn.org/stable/
https://github.com/monsterautomata/text-klasifikasi-gender