Laporan Klasifikasi Teks Opini Terhadap Prabowo di Media Sosial Twitter ke dalam Positif dan Negatif

Tema : Klasifikasi Sentimen

Ditunjukan untuk memenuhi tugas 3 Pemrosesan Bahasa Alami terkait *Klasifikasi Teks*.

Oleh :

Alfian Yulianto – 1301178160

Akbar Habib Buana Wibawa Putra – 1301178198

Bastomy – 1301178418

Muhammad Hanafiah – 1301178552

ICM-39-GAB

S1 Teknik Informatika

Fakultas Informatika

Telkom University

2018

A. Abstrak

Masa pemilihan presiden merupakan masa dimana masyarakat menyampaikan opini-opini mereka terhadap para calon presiden selanjutnya, dengan adanya sosial media para masayarakat semakin mudah dalam menyampaikan opini mereka, mulai dari pendukung paslon itu sendiri hingga kubu opisisi saling mengomentari para calon presiden, salah satu contohnya pada media sosial twitter yang banyak sekali kicauan terhadap para calon presiden yang dimana terdapat banyak kicauan mulai dari yang baik hingga menjelekkan calon presiden yang ada, melalui tugas besar ini akan dilakukan klasifikasi berdasarkan kicauan para netizen terhadap salah satu calon presiden pada media sosial twitter dengan menggunakan metode naive bayes classifier untuk mendapatkan hasil akurasi pengklasifikasian.

B. Pendahuluan

Pada tugas ini akan dilakukan pengklasifikasian terhadapa kicauan pada media sosial twitter terhadap salah satu calo presiden yakni bapak Prabowo, klasifikasi ini dilakukan untuk mengetahui komentar-komentar yang mengarah pada calon presiden Prabowo apakah bersifat positif atau negatif. pada penyusunan tugas ini akan dilakukan secara bertahap yang dimulai dari pengumpulan data set yang didapatkan dari crawling data media sosial twitter minimal 100 data, selanjutnya data akan dipraproses agar lebih untuk mendapatkan data yang lebih bersih agar hasil akurasi lebih baik, selanjutnya penerapan metode Naive bayes untuk membangun classifier yang diterapkan kepada komentar netizen untuk mengetahui apakah komentar mereka bersifat positif atau negatif.

C. Keterangan Korpus

korpus yang digunakan adalah data tweet yang berkaitan dengan kata kunci "prabowo" yang dimana menjadi objek utama dari tugas ini. data di crawling langsung dengan

menggunakan API twitter dengan menggunakan bahasa python, data disimpan dalam bentuk csv. data akan dilabeli terlebih dahulu untuk menjadi data training nantinya.

D. Tema : Klasifikasi Sentimen

Pada tugas ini termasuk dalam kategori sentimen, yaitu postif negatif dari komentar yang ada pada media sosial twitter terhadap bapak Prabowo

E. Fitur yang dipilih

Fitur yang dipilih dalam pembangunan system sebagai berikut :

- Crawling: pengambilan 100 data tweet dari media sosial twitter dengan menggunakan Bahasa pemograman phyton dan API twitter agar dapat mengakses data-data tweet.
- Labelling: pemberian label dari data set yang telah dikumpulkan, terdapat 2 jenis label yaitu label positif dan label negatif dari suatu tweet.
- Preprocessing : preprocessing adalah tahap mempersiapkan data sebelum di proses, dimana pada tahap preprocessing ini terdapat beberapa poses yaitu
    - Casefold: merubah kalimat tweet menjadi lowercase
    - Tokenisasi: merubah kalimat tweet menjadi sebuah token
    - stopword removal: menghapus kata yang tidak memiliki makna
    - remove symbol dan character: menghapus symbol, link, mention dan hashtag.
- Naïve Bayes: sebuah metode klasifikasi probabilitas sederhana yang mengaplikasikan Teorema Bayes dengan asumsi ketidak tergantungan yang tinggi.
- Laplace Smoothing: smoothing adalah suatu cara untuk menangani nilai probabilitas 0.
- cross validation: suatu Teknik pengujian yang digunakan dengan cara melakukan looping dengan mengacak atribut masukan.

F. Teknik yang dipakai

a. Naïve Bayes

Naive Bayes Classifier merupakan sebuah metode klasifikasi probabilitas sederhana yang mengaplikasikan Teorema Bayes dengan asumsi ketidak tergantungan (independent) yang tinggi. Algoritma *Naïve Bayes Classifier* memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya. Proses yang terdapat pada algoritma ini ada dua tahap, yaitu proses pelatihan (*learning*) dan proses klasifikasi.

Untuk setiap dokumen, *Naïve Bayes Classifier* menghitung *posterior probability* dimana dokumen termasuk pada kelas yang berbeda. Dokumen tersebut dikelompokan kedalam kelas yang memiliki nilai *posterior probability* paling besar. Secara matematis, NBC dituliskan sebagai berikut [1] :

$$P(A|B) = \frac{\left(P(B|A) \times P(A)\right)}{P(B)}$$

Dimana:

- A = sampel data yang label kelasnya tidak diketahui.
- B = kelas-kelas hasil klasifikasi.
- P(A|B) = probabilitas terjadinya A jika B diketahui. Disebut probabilitas *posterior*, karena peluang A bergantung dari nilai B tertentu.
- P(B|A) = probabilitas terjadinya B jika A diketahui, disebut *likelihood* function, karena peluang B tergantung dengan peluang data sample A.

- P(A) = probabilitas A merupakan probabilitas dari sample yang mempunya kelas A.

  P(B) = probabilitas prior B, dan bertindak sebagai *normalizing constant*.
  Secara intuitif , teorema Bayes menggambarkan bahwa perubahan pada "A" dapat diamati apabila "B" terlebih dahulu diamati

b. Laplace Correction

Laplace Correction (Laplacian Estimator) atau additive smoothing adalah suatu cara untuk menangani nilai probabilitas 0 (nol). Dari sekian banyak data di training set, pada setiap perhitungan datanya ditambah 1 (satu) dan tidak akan membuat perbedaan yang berarti pada estimasi probabilitas sehingga bisa menghindari kasus nilai probabilitas 0 (nol).

$$Pi = \frac{Mi + 1}{n + k}$$

dimana nilai k adalah jumlah kelas atau bin dari atribut mi.

Sebagai contoh, asumsikan ada class buy=yes disuatu training set, memiliki 1000 sampel, ada 0 (nol) sampel dengan income=low, 990 sampel dengan income=medium, dan 10 sampel dengan income=high. Probabilitas dari kejadian ini tanpa Laplacian Correction adalah 0, 0.990 (dari 990/1000), dan 0.010 (dari 10/1000). Menggunakan Laplacian Correction dari tiga sampel diatas, diasumsikan ada 1 sampel lagi untuk masing – masing nilai income. Dengan cara ini, didapatkanlah probabilitas sebagai berikut (dibulatkan menjadi 3 angka dibelakang koma):

1/1003 = 0.001 | 991/1003 = 0.988 | 11/1003 = 0.011

Probabilitas yang "dibenarkan" hasilnya tidak berbeda jauh dengan hasil probabilitas sebelumnya sehingga nilai probabilitas 0 (nol) dapat dihindari [2].

G. Teknik Pengujian

a. K-Fold Cross Validation

K-Fold cross validation adalah salah satu metode yang digunakan untuk mencari tahu rata-rata keberhasilan dari suatu sistem yang dibangun dengan cara melakukan *looping* dengan mengacak atribut masukan sehingga suatu sistem teruji untuk beberapa atribut input yang acak. *K-Fold cross validation* dimulai dengan membagi data sejumlah *nfold* yang diinginkan. Dalam proses *cross validation* nantinya data akan dibagi dalam n buah partisi yang memiliki ukuran yang sama D1, D2, D3. Dn selanjutnya proses *testing* dan *training* dilakukan sebanyak n kali. Dalam iterasi ke-i partisi Di akan menjadi data *testing* dan selebihnya akan menjadi data *training*. Untuk penggunaan jumlah *fold* terbaik untuk uji validitas, disarankan menggunakan *10-Fold Cross Validation* dalam model. Skenario pengujian merupakan tahap penentuan pengujian yang akan dilakukan. Pengujian akan dilakukan menggunakan metode *k-cross validation* dengan nilai k sebanyak *10 Fold*, pengujian ini memiliki tujuan untuk mendapatkan akurasi metode *Naive Bayes* yang diimplementasikan pada analisis *spam* jika diuji dengan data *training* dan data *testing* yang berbeda. Penggunaan *10 Fold* ini dianjurkan karena merupakan jumlah *Fold* terbaik untuk uji validitas [3].

Metode yang digunakan adalah metode *10 fold cross validation* yaitu *dataset* akan dibagi menjadi 10 bagian sama rata. Pada *fold* pertama, terdapat kombinasi 9 *subset* yang berbeda digabungkan dan digunakan sebagai *data training*, sedangkan satu bagian digunakan untuk *data testing*, selanjutnya proses *training* dan *testing* dilakukan hingga *fold* kesepuluh. Skenario *10 fold cross validation* data dilihat pada table berikut:

Fold	Data Training	Data Testing
1	D2,D3,D4,D5,D6,D7,D8,D9,D10	D1
2	D1,D3,D4,D5,D6,D7,D8,D9,D10	D2
3	D1,D2,D4,D5,D6,D7,D8,D9,D10	D3
4	D1,D2,D3,D5,D6,D7,D8,D9,D10	D4

Fold	Data Training	Data Testing
5	D1,D2,D3,D4,D6,D7,D8,D9,D10	D5
6	D1,D2,D3,D4,D5,D7,D8,D9,D10	D6
7	D1,D2,D3,D4,D5,D6,D8,D9,D10	D7
8	D1,D2,D3,D4,D5,D6,D7,D9,D10	D8
9	D1,D2,D3,D4,D5,D6,D7,D8,D10	D9
10	D1,D2,D3,D4,D5,D6,D7,D8,D9	D10

b. Skenario Pengujian Naïve Bayes

Berikut adalah langkah-langkah pengujian yang dilakukan dengan beberapa contoh.

1. ***Tentukan data latih dan data tes dalam data set***

No	Tweet	Class
1	2019 ganti presiden	Negatif
2	Pelemahan rupiah di pemerintahan Jokowi	Negatif
3	Kerja nyata pembangunan bendungan	Positif
4	Pemerintahan jokowi pembangunan tol semakin banyak	?

2. ***Ubah data set kedalam frekuensi data***

Kata	Positif	negatif
Ganti	0	1
Presiden	0	1
Jokowi	1	1
Pelemahan	0	1
Rupiah	0	1
Kerja	1	0
Nyata	1	0
Pembangungan	1	0
bendungan	1	0
Pemerintahan	1	1

3. *Hitung Prior*

P(positif) = 1/3 = 0.33

P(negatif) = 2/3 = 0.64

$$P(C) = \frac{N_C}{N}$$

Dimana :

P(C) adalah probabilitas dari kelas

Nc adalah jumlah total kelas tertentu di data latih

N total kelas di data latih

4. *Hitung probabilitas bersyarat / likelihood setiap kata (1)*

$$P(w|c) = \frac{count(w,c) + 1}{count(C) + |V|}$$

Dimana :

P(w|c) adalah kondisi likelihood

W adalah atribut kata

C adalah kelas

+1 untuk menghindari pembagian dengan 0

Count(c) total kata di kelas tertentu

|V| total kata

5. *Hitung probabilitas bersyarat / likelihood setiap kata (2)*

P(pemerintah | positif)	1+1 / 6+10 = 2/16	0.12
P(Jokowi | positif)	1+1 / 6+10 = 2/16	0.12
P(pembangunan | positif)	1+1 / 6+10 = 2/16	0.12
P(tol | positif)	0+1 / 6+10 = 1/16	0.05
P(pemerintah | negatif)	1+1 / 6+10 = 2/16	0.12
P(Jokowi | negatif)	1+1 / 6+10 = 2/16	0.12
P(pembangunan | negatif)	0+1 / 6+10 = 1/16	0.05
P(tol | negatif)	0+1 / 6+10 = 1/16	0.05

6. *Hitung probabilitas posterior*

• P(positif) = P(pemerintah|positif) * P(Jokowi|positif) * P(pembangunan|positif) * P(tol|positif)*P(positif)

= (0.12) * (0.12) * (0.12) * (0.05) * (0.33)

= 0.000028512

• P(negatif) = P(pemerintah|negatif) * P(Jokowi|negatif) * P(pembangunan|negatif) * P(tol|negatif)*P(negatif)

= (0.12) * (0.12) * (0.0.5) * (0.05) * (0.64)

= 0.00002304

7. *Tentukan kelas dari data tes*

P(positif) = 0.000028512

P(negatif) = 0.00002304

Ambil nilai terbesar, maka "Pemerintahan jokowi pembangunan tol semakin banyak" = Positif

H. Evaluasi dan Analisis Hasil Klasifikasi

```
f1,accuracy=0,0
for i in range(len(result)):
    f1 = f1+result['F1'][i]
    accuracy=accuracy+ result['accuracy'][i]
print("rata-rata F1 ",f1/len(result))
print("rata-rata accuracy ",accuracy/len(result))
```

```
rata-rata F1  0.6861236802413272
rata-rata accuracy  0.6262626262626262
```

Setelah dilakukan pengujian pengklasifikasian pada tugas ini diperoleh hasil akhir F1 bernilai sekitar 0,687 dan rata-rata akurasi 0.627 yang dimana bisa dikatakan nilainya kurang dari rata-rata yang cukup bagus, karena rata-rata akurasi yang bisa dikatakan mulai cukup bagus hingga bagus yaitu sekitar mulai dari 0.7 hingga 0.8, disini dataset yang digunakan juga sangat mempengaruhi pada saat *cross validation* dan perhitungan akurasinya.

I. Kesimpulan

Berdasarkan hasil dari penelitian dan analisis yang dilakukan dapat dikatakan bahwa metode *naive bayes* yang digabungkan dengan fitur-fitur yang diimplementasikan dapat melakukan klasifikasi *tweet* sesuai dengan kelas/label yang telah ditentukan, dengan catatan penting pra proses yang dilakukan benar sesuai kebutuhan dan data training yang dimiliki mumpuni dan cukup relevan sehingga akurasi yang diperoleh nantinya akan jadi lebih besar, disini kami mendapatkan presentase sekitar 63%. Dari hasil presentase yang didapatkan bisa dikatakan tantangan dalam pembuatan tugas besar ini adalah ketika proses pencarian dataset yang sesuai dengan kriteria yang dibutuhkan serta menentukan proses preprocessing dan fitur yang dibutuhkan untuk mendapatkan hasil klasifikasi yang sesuai.

J. Referensi

1. S. F. Rodiyansah and E. Winarko, "Klasifikasi Postingan Twitter Kemacetan Lalu Lintas Kota Bandung Menggunakan Naive Bayessian Classification," Yogyakarta, Universitas Gajah Mada, 2014.
2. Online https://informatikalogi.com/algoritma-naive-bayes/. Diakses 11 mei 2019.
3. Kohavi, "A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," 2014, p. 167.