

Teks Mining Untuk Klasifikasi Keluhan Atlet Pada Koni Kota Probolinggo

Dyah Ariyanti

Program Studi Teknik Elektro, Fakultas Teknik, Universitas Panca Marga Probolinggo

diayantiku@gmail.com

Terima Naskah : 3 April 2018

Terima Revisi : 17 April 2018

ABSTRAK

Penanganan keluhan / pengaduan atlet Kota Probolinggo masih memerlukan waktu yang lama untuk dapat ditangani oleh Koni dikarenakan seorang admin terkadang tidak mengetahui harus kemana keluhan atlet tersebut ditujukan. Penelitian tentang teks mining untuk klasifikasi keluhan masyarakat dengan algoritma naïve bayes diharapkan dapat membantu admin dalam memecahkan persoalan diatas, sehingga waktu untuk menjawab keluhan atlet tersebut menjadi lebih efektif dan efisien. Pengolahan data keluhan atlet ini melalui beberapa tahapan teks mining yaitu *token*, *filter*, *stemming* dan *analyzing*. Setelah melalui semua tahapan ini, data tersebut akan diklasifikasikan menggunakan algoritma *Naïve Bayes*. Hasil perhitungan tersebut nantinya akan menunjukkan hasil kelas setiap data keluhan atlet yang masuk, melalui kuisioner maupun email. **Kata kunci** : *Text mining, Token, Filter, Stemming, Analyzing, Naïve Bayes*.

ABSTRACT

Handling athlete complaints in Probolinggo city, requires more time to reach the relevant by Koni it is because the administrators sometimes do not know where to address the athlete complaints. Using naïve bayes algorithm in text mining, it is expected that these public complaints in Probolinggo can help the administrators work more effectively and efficiently. The data processing has several stages of text mining : token, filter, steaming and analyzing. After completed these stages, the data will be classified using naïve bayes algorithm. The result of the calculation with this algorithm will show the classes of athlete complaints in Probolinggo received through questioner and email.

Kata kunci : *Text mining, Token, Filter, Stemming, Analyzing, Naïve Bayes*.

PENDAHULUAN

Koni Kota Probolinggo saat ini sedang menerapkan informasi keterbukaan kepada seluruh lapisan masyarakat khususnya pecinta olahraga di Kota Probolinggo. Informasi yang diberikan tentunya tidak terbatas dan terbuka untuk segala aspek.

Seiring perkembangan teknologi dan ilmu pengetahuan tersebut, maka keluhan yang disampaikan semakin berkembang dan menjadi semakin luas, tidak hanya melalui catatandan melalui media email.

Data yang didapatkan / diperoleh semakin banyak dan beragam. Hal ini sangat mempengaruhi efisiensi dan kecepatan jawaban seperti yang diharapkan oleh atlet Kota Probolinggo.

Oleh karena itu dibutuhkan suatu sistem sebagai alternatif lain untuk memperoleh jawaban yang akurat dan tidak lagi membutuhkan waktu yang lama dalam merespon. Salah satu algoritma yang dapat digunakan untuk mengklasifikasikan dokumen dimulai dari pengelolaan dokumen teks menggunakan *text mining*.

Dalam proses penentuan kelas untuk Cabang Olahraga (Cabor) dapat menggunakan algoritma

Naïve Bayes (NB). Pada kasus ini menggunakan *sample data training* sebanyak 100 data keluhan atlet dan *data testing* sebanyak 50 data keluhan atlet. Rumusan masalah yang dibahas pada penelitian ini yaitu bagaimana menerapkan *text mining* untuk klasifikasi keluhan masyarakat menggunakan algoritma NB.

Penentuan kelas dari Cabor Kota Probolinggo sebanyak 10 (sepuluh) Cabor. Dengan menerapkan *Teks Mining*, selanjutnya proses awal yang dilakukan adalah memasukkan data *training* yang sudah dikelompokkan secara manual dan menentukan *keyword / kategori* yang bervariasi. Dari beberapa *keyword* apabila tidak sesuai dengan penentuan kelas nantinya bisa dimasukkan di dalam daftar *Stopword*. Data kalimat keluhan masyarakat yang digunakan untuk penentuan *keyword* tersebut dimasukkan sebanyak 100 (seratus) Cabor yang dipilih, baik dari telepon, kuisioner dan website. Apabila pra proses telah terpenuhi selanjutnya dilakukan perhitungan probabilitas NB pada tiap-tiap *keyword*, sehingga akhirnya siap untuk dilakukan pengujian menggunakan data *testing*.

Tujuan dan manfaat dari penelitian yang dilakukan adalah :

a. Membuat sebuah sistem pengelompokan kategori / keyword data keluhan masyarakat berdasarkan seluruh keluhan masyarakat yang tersedia pada KONI Kota Probolinggo dengan menerapkan Algoritma NB.

b. Dengan menerapkan algoritma NB melakukan proses pengklasifikasian text dari berbagai media yang tersedia sehingga nantinya dapat menghasilkan informasi secara tepat dan akurat dari keluhan atlet untuk cabor terkait

METODE

*Teks mining*¹ adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Jadi, sumber data yang digunakan pada *teks mining* adalah kumpulan teks yang memiliki format yang tidak terstruktur atau minimal semi terstruktur. Adapun tugas khusus dari teks mining antara lain yaitu pengkategorisasian teks (*teks categorization*) dan pengelompokan teks (*teks clustering*).

Permasalahan yang dihadapi pada teks mining sama dengan permasalahan yang terdapat pada data mining, yaitu jumlah data yang besar, dimensi yang tinggi, data dan struktur yang terus berubah, dan data *noise*.

Probabilitas Bayesian² adalah suatu interpretasi dari kalkulus yang memuat konsep probabilitas sebagai derajat dimana suatu pernyataan dipercaya benar. Teori Bayesian juga dapat digunakan sebagai alat pengambilan keputusan untuk memperbaharui tingkat kepercayaan dari suatu informasi. Prediksi Naïve Bayes, didasarkan pada teorema Bayes dengan formula umum sebagai berikut :

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

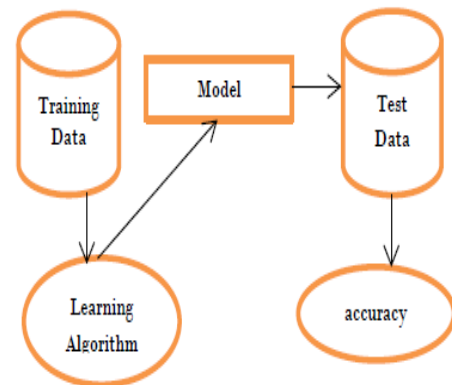
Penjelasan :

- P(A|B) = Probabilitas akhir bersyarat (conditional probability) suatu hipotesis A terjadi jika diberikan bukti evidence (B) terjadi.
- P(B|A) = Probabilitas sebuah bukti B terjadi akan mempengaruhi hipotesis A.
- P(A) = Probabilitas awal (priori) hipotesis A terjadi tanpa memandang bukti.
- P(B) = Probabilitas awal (priori) bukti B terjadi tanpa memandang bukti yang lain.

Teorema Bayes menerangkan hubungan antara probabilitas terjadinya peristiwa A dengan syarat peristiwa B telah terjadi dan probabilitas terjadinya peristiwa B dengan syarat peristiwa A telah terjadi. Tambahan informasi dapat memperbaiki probabilitas.

HASIL DAN PEMBAHASAN

Berikut ini adalah ilustrasi sistem yang menunjukkan alur kerja pada aplikasi :

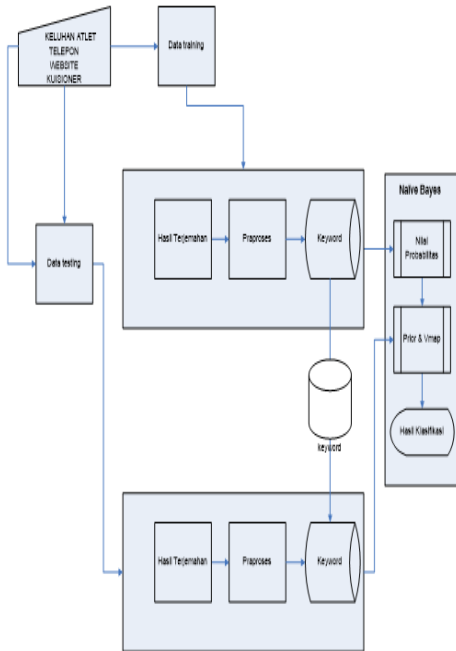


Gambar 1. Ilustrasi sistem

Kontribusi yang diharapkan dari penelitian ini adalah sistem mampu memberikan hasil pengelompokan Cabor secara efektif dan efisien. Sehingga penyampaian keluhan atlet ke tiap-tiap Cabor cepat untuk direspon, dan tentunya atlet Kota Probolinggo segera mungkin untuk memperoleh jawaban yang akurat dan tidak membutuhkan waktu yang lama.

Proses awal pertama kali yang dilakukan pada penelitian ini adalah melakukan testing data keluhan atlet yang sudah dikelompokkan secara manual oleh KONI Kota Probolinggo, yang diterima dari email ataupun kuisioner. Data yang diinputkan sebanyak 100 (seratus) komentar, yang dipilih dari 10 (sepuluh) Cabor.

Bagian ini akan menjelaskan beberapa proses sistem yang akan dijalankan pada aplikasi penelitian ini.



Gambar 2. Beberapa proses pada sistem

tersimpan, selanjutnya akan menentukan dalam perhitungan probabilitas. Proses berikutnya adalah melakukan perhitungan probabilitas naïve bayes dengan menggunakan persamaan rumus :

$$P(x_i|v_j) = \frac{n_k + 1}{n + |\text{kosakata}|}$$

$$P(v_j) = \frac{|\text{docs } j|}{|\text{contoh}|}$$

Keterangan :

- |docs j| = jumlah dokumen setiap kategori j
- |contoh| = jumlah dokumen dari semua kategori
- nk = jumlah frekuensi kemunculan setiap kata,
- n = jumlah frekuensi kemunculan kata dari setiap kategori,
- |kosakata| = jumlah semua kata dari semua kategori.

Dari hasil perhitungan menggunakan rumus tersebut, maka akan dapat menentukan hasil akhir kelas Keluhan_Atlet dengan rumus dibawah ini.

$$v_{MAP} = \underset{v \in V}{\text{arg max}} \prod_{i=1}^n (P(x_i|V_j)P(v_j))$$

Berikut ini adalah salah satu contoh bentuk uji coba untuk membuktikan teks mining klasifikasi keluhan masyarakat dengan menggunakan algoritma NB. Contoh Testing Keluhan Atlet :

- Setiap bulan latihan bola voli kami tidak pernah menerima gizi?

Token

tiap // bulan // latihan // bola // voli // kami // tidak // pernah // terima // gizi

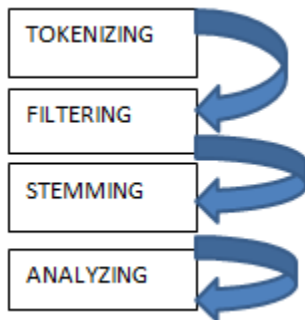
Hasil Stemming – Stop Word voli (1), gizi (1) Selanjutnya diambil perhitungan nilai probabilitas yang sesuai dengan keyword yang termasuk pada contoh tersebut. Nilai probabilitas keyword yang sesuai. Dan

kemudian dilakukan perhitungan klasifikasi dengan yang hasilnya dijabarkan berikut ini :

Tabel 1. hasil perhitungan klasifikasi

Kelas	voli	gizi	P(Ci)	VMAP
PSSI	0.0044646	0.004	0.1	0.0000019

Data yang dimasukkan tersebut selanjutnya dilakukan proses teks mining³.



Gambar 3. Proses teks mining

Dari hasil akhir proses stemming kemudian selanjutnya komentar tersebut dibuatkan daftar stop list. Kata yang tidak termasuk dalam stop list, akhirnya disimpan dalam database keyword yang telah disiapkan. Apabila daftar keyword tersebut termasuk kata umum yang sering digunakan pada semua cabor, maka bisa dipilih untuk dihapus (dimasukkan dalam daftar stop list). Sehingga akan terlihat jumlah kata yang

PERBASI	0.00412	0.004	0.1	0.0000016 93509
PBSI	0.00382	0.003	0.1	0.0000014 56792
PSTI	0.00388	0.003	0.1	0.0000015 02314
PERSANI	0.00403	0.004	0.1	0.0000016 25911
PBVSİ	0.03543	0.015	0.1	0.0000558 00112
TI	0.00376	0.003	0.1	0.0000014 13308
PERCASI	0.00364	0.003	0.1	0.0000013 22314
IPSI	0.00411	0.016	0.1	0.0000067 74035
FORKI	0.00386	0.003	0.1	0.0000014 90735

P(“PBVSİ”|”Keluhan_Atlet”) mempunyai nilai yang lebih besar dibandingkan dengan 9 (sembilan) kelaslainnya sehingga hasil data training pada tabel 1 masuk pada kelas PBVSİ.

Segmen Program Menghitung Naïve Bayes

```

:
1 : baris = 0
2 : VSNB.Clear
3 : For kolom = 1 To GridNB.Cols - 1

4 : nilaiawal = 0

5 : nilaiakhir = 1
6 : For j = 1 To GridNB.rows - 1
7 : For i = 1 To VSDocMuncul.rows
- 1
8 : If VSDocMuncul.TextMatrix(i,
1) = GridNB.TextMatrix(j, 0) Then
9 : baris = baris + 1

10 : VSNB.rows = baris + 1
11 : VSNB.TextMatrix(baris, 0) = baris
12 : VSNB.TextMatrix(baris, 1) =
"P ( " & GridNB.TextMatrix(0,
kolom) & " | " &
GridNB.TextMatrix(j, 0) & " )

```

```

= " & GridNB.TextMatrix(j, kolom) &
""
13 : VSNB.TextMatrix(baris, 2) =
GridNB.TextMatrix(j, kolom) 'hasil
14 : VSNB.TextMatrix(baris, 3) =
GridNB.TextMatrix(0, kolom) 'dinas
15 : VSNB.TextMatrix(baris, 4) =
GridNB.TextMatrix(j, 0)
'keyword
16 : nilaiawal =
CDBl(GridNB.TextMatrix(j, kolom))
17 : nilaiakhir = nilaiakhir * nilaiawal
18 : End If
19 : Next i

20 : Next j
21 : nilaiawal = GridCabor.rows -
1
22 : For i = 1 To GridCabor.rows -
1
23 : If GridCabor.TextMatrix(i, 2)
=
GridNB.TextMatrix(0, kolom)
Then
24 : nilaiakhir = (1 / nilaiawal)
*
nilaiakhir '( 1/totalcabor)
*
nilai probabilitas per kosakata
25 : GridCabor.TextMatrix(i, 9) =
CDBl(Format(nilaiakhir,
"##.#####"))
26 : End If

27 : Next i
28 : Next kolom

29 : nilaiawal = GridCabor.TextMatrix(1,
9) 'catch nilai dinas pertama

30 : nilaiakhir = GridCabor.rows - 1 'catch
maksimal indeks grid Cabor
31 : For i = 1 To GridCabor.rows - 1

32 : If nilaiawal =
Val(GridCabor.TextMatrix(i, 9)) Then

```

```

33 : kalimat = GridCabor.TextMatrix(i, 2)

34 : xkdcabor = GridCabor.TextMatrix(i, 8)

35 : ElseIf Val(GridCabor.TextMatrix(i,
9)) > nilaiawal Then
36 : nilaiawal =
Val(GridCabor.TextMatrix(i, 9))

37 : kalimat = GridCabor.TextMatrix(i, 2)
38 : xkddinas = GridCabor.TextMatrix(i,
8)
39 : End If

40 : Next i

```

SIMPULAN

Dari hasil uji coba pada penelitian ini maka dapat ditarik kesimpulan sebagai berikut :

1. Hasil Klasifikasi yang diperoleh dalam penelitian ini mendapatkan hasil yang baik. Uji coba yang pertama dilakukannya uji klasifikasi dengan menggunakan semua proses sampai tahap analyzing data mendapatkan hasil nilai rata-rata akurasi sebesar 88 %.
2. Hasil terbaik yang diberikan selama uji coba pada penelitian ini adalah melakukan klasifikasi terhadap keluhan masyarakat, dimana data yang dipakai bersumber dari kuisisioner dan email.

DAFTAR PUSTAKA

- [1] https://en.wikipedia.org/wiki/Text_mining
- [2] http://en.wikipedia.org/wiki/Bayesian_probability
http://en.wikipedia.org/wiki/Naive_Bayes_classifier
- [3] Triawati Candra. (2009). *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia*.SNATI STT Telkom: Bandung.