unplug Classification of Student Complaints with Naive Bayes and Literary Methods

by Mochamad Alfan Rosid

Submission date: 09-Jan-2024 10:42AM (UTC+0700)
Submission ID: 2268214251
File name: 711-Article_Text-9537-3-10-20220810.pdf (794.87K)
Word count: 2351
Character count: 14251

Classification of Student Complaints with Naive Bayes and Literary Methods

Klasifikasi Keluhan Mahasiswa dengan Metode Naive Bayes dan Sastrawi

Haris Ahmad Gozali, Mochamad Alfan Rosid*, Sumarno

Informatika, Universitas Muhammadiyah Sidoarjo, Indonesia

Email Penulis Korespondensi: alfanrosid@umsida.ac.id

Abstract. *E-Complaint Data is a collection of data containing comments or complaints from students against the University. Many comments: comments that occur in the university environment about the facilities for the performance of teachers, etc. Text mining, also known as text data mining or search for knowledge in textual databases, is a semi-automatic process for extracting patterns from data. The purpose of text mining is to obtain useful information from a set of documents. This study uses the Bayes ship method with the TFIDF weighting function. The stages that will be taken to determine the ranking. The first is to take data from the E-Complaint System, then the data will go through the preprocessing stage using a literary library, after going through the preprocessing stage, the data will be divided into 2, namely, training data and data from proof. Then the training data will be carried out using the TF-IDF weighting process to probability, if so, the next step is to process the test data by determining the previous values. The next is the data test stage between the test data and the training data, then the data test results will be in the form of default categories. The test results show that the classification of complaints with the Bayes ship algorithm and with the TF-IDF function and the literary library in the preprocessing process has a fairly high average precision of 82%.*

Keywords- *Classification, Student Complaint, Naïve Bayes, Sastrawi*

Abstrak. *Data Pengaduan Elektronik adalah kumpulan data yang berisi komentar atau pengaduan dari mahasiswa terhadap Universitas. Banyak komentar: komentar yang terjadi di lingkungan universitas tentang fasilitas untuk kinerja guru, dll. Text mining, juga dikenal sebagai text data mining atau pencarian pengetahuan di database tekstual, adalah proses semi-otomatis untuk mengekstraksi pola dari data. Tujuan dari text mining adalah untuk mendapatkan informasi yang berguna dari sekumpulan dokumen. Penelitian ini menggunakan metode kapal Bayes dengan fungsi pembobotan TFIDF. Tahapan yang akan dilakukan untuk menentukan ranking. Yang pertama adalah mengambil data dari E-Complaint System, kemudian data tersebut akan melalui tahap preprocessing menggunakan literature library, setelah melalui tahap preprocessing data akan dibagi menjadi 2 yaitu data training dan data dari proof . Kemudian akan dilakukan data latih dengan menggunakan proses pembobotan TF-IDF terhadap probabilitas, jika sudah maka langkah selanjutnya adalah mengolah data uji dengan menentukan nilai sebelumnya. Selanjutnya adalah tahap uji data antara data uji dan data latih, kemudian data hasil uji akan berupa kategori default. Hasil pengujian menunjukkan bahwa klasifikasi komplain dengan algoritma kapal Bayes dan dengan fungsi TF-IDF dan library literature pada proses preprocessing memiliki rata-rata presisi yang cukup tinggi yaitu 82%.*

Kata kunci- *Klasifikasi, Keluhan Mahasiswa, Naïve Bayes, Sastrawi*

PENDAHULUAN

Penambangan teks (text mining) adalah metode untuk mengekstrak informasi yang berguna dari teks dengan bantuan program komputer. Istilah informasi yang berguna dapat merujuk ke berbagai informasi tergantung pada tujuan penggunaan metode tersebut. Klasifikasi teks, pengelompokan, ekstraksi nilai numerik, peringkasan dokumen adalah beberapa dari banyak sub-cabang metode. Metode penambangan teks yang tidak biasa menggunakan satu atau kombinasi dari analisis pola statistik, analisis leksikal, analisis frekuensi kemunculan, penandaan, dll[1][2][3].

Dalam penelitian Mochamad Alfan Rosid dkk, klasifikasi keluhan mahasiswa dilakukan dengan menggunakan metode classifier berbasis centroid dan dengan fitur TF-IDF-ICF, selain itu pada proses stemming menggunakan metode porter Z Tala. Pada sistem informasi di atas, data diambil dari aplikasi klaim elektronik Universitas Muhammadiyah Sidoarjo. Hasil percobaan memperlihatkan bahwa klasifikasi keluhan dengan metode classifier berbasis centroid dan dengan fungsi TF-IDF-ICF memiliki rata-rata akurasi 79,5% [4].

TF-IDF merupakan gabungan dari dua metode untuk melakukan pembobotan kata, yaitu frekuensi kemunculan suatu kata dalam dokumen tertentu dan frekuensi kebalikan dari dokumen yang memiliki kata tersebut. Frekuensi kemunculan sebuah kata dalam dokumen tertentu menunjukkan pentingnya kata tersebut dalam dokumen. Frekuensi dokumen yang berisi kata tersebut memperlihatkan seberapa umum kata tersebut. Jadi bobot hubungan antara kata dan dokumen akan tinggi jika

frekuensi kata tinggi dalam dokumen dan frekuensi seluruh dokumen yang mengandung kata rendah dalam kumpulan dokumen. Selain pembobotan menggunakan TF-IDF, peneliti juga menggunakan tahap preprocessing, sedangkan tahapan yang digunakan adalah stopword removal dan derivation.[5].

Tahapan yang digunakan adalah stopwords removal, yaitu proses menghilangkan kata-kata yang tidak penting dalam deskripsi dengan cara mengecek kata-kata yang ada di deskripsi parse, apakah termasuk dalam daftar kata tidak penting (stoplist) atau tidak. Kedua adalah tahap stemming, yaitu tahap untuk mengubah token menjadi token berupa kata-kata dasar. Tujuan dari tahap stemming adalah untuk mengurangi jumlah kata yang memiliki kata dasar yang sama[6].

Pada penelitian ini mencoba mengusulkan penggunaan metode naïve bayes dalam tahap klasifikasi dan mencoba menggunakan library sastrasi dalam proses mengubah kata imbuhan menjadi kata dasar. Tujuan dari penelitian ini adalah untuk meningkatkan akurasi klasifikasi dari penelitian yang sudah dilakukan oleh Mochamad Alfan Rosid dkk.

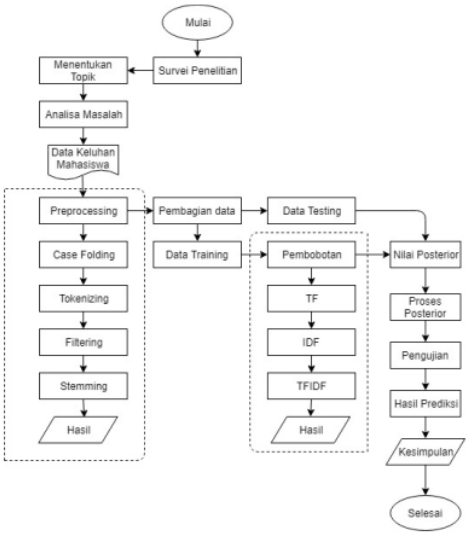
Algoritma Naive Bayes adalah salah satu algoritma terdapat pada teknik klasifikasi. Metode klasifikasi Naive Bayes adalah probabilitas dan statistik yang dimunculkan oleh ilmuwan Inggris Thomas Bayes, yang memprediksi peluang masa depan berdasarkan pengalaman sebelumnya dan kemudian dikenal sebagai Teorema Bayes. Teorema naif digabungkan dengan kondisi atribut yang diasumsikan independen. Klasifikasi Naive Bayes diasumsikan bahwa ada atau tidaknya karakteristik tertentu dari suatu kelas tidak ada hubungannya dengan karakteristik kelas lainnya.[7]

METODE PENELITIAN

Dalam melaksanakan penelitian ini, kami melakukan beberapa tahap, yaitu dimulai dari pengambilan data, tahap preprocessing, pembagian data, pembobotan, penerapan metode naïve bayes sampai data berhasil terklasifikasi. Adapun tahapan penelitian tersebut dapat digambarkan oleh Gambar 1.

Dataset yang digunakan berjumlah 573 yang berasal dari data komentar/keluhan mahasiswa di aplikasi E-Complaint. Dataset ini meliputi katagori sebagai berikut :

- Kemahasiswaan
- DPAL (Direktorat Pengelolaan Aset dan Lingkungan)
- DA (Direktorat Akademik)
- DK (DIrektorat Keuangan)
- DSTI (Direktorat Sistem dan Teknologi)
- PERPUSTAKAAN
- FAKULTAS
- DRPM (Direktorat Riset dan Pengabdian Masyarakat)
- LAIN-LAIN
- BPM (Badan Penjaminan Mutu)



Gambar 1. *Tahapan Penelitian*

Data keluhan mahasiswa, ditunjukkan oleh Tabel 1.

Tabel 1. *Contoh data keluhan mahasiswa*

Dokumen Keluhan	Kategori
UKM agar difasilitasi computer dan printer	Kemahasiswaan
Assalamualaikum Pengadaan inventaris alat musik tolong bantuaannya karna alat musik itu kebutuhan pokok buat UKM IKABAMA	Kemahasiswaan
Pak/bu seketariat UKM IKABAMA tidak ada studio beserta sebagian alat musiknya.	Kemahasiswaan
LCD/proyektor diperbaiki , gambarnya sudah tidak jelas di ruang 401 gedung C	DPAL
Pintu kelas 201 diperbaiki daun pintunya sudah putus jadi sulit buka pintunya	DPAL
Disediakan sound/speaker pada tiap tempat pelayanan agar terdengar kalau ada yang dipanggil	DPAL

Tahap Prepocessing

Tahap Preprocessing disini adalah pembersihan data, fungsi dari tahap ini adalah agar akurasi yang didapat menjadi baik. Tahap Preprocessing terdiri dari 4 bagian[8][9], yaitu :

a. Tahap Case Folding : Tahap ini merupakan proses mengubah seluruh kalimat menjadi huruf kecil

b. Tahap Tokenizing : Tahap ini adalah sebuah proses

penguraian deskripsi yang semula berupa kalimat menjadi sebuah kata

c. Tahap Stemming : Tahap ini merupakan proses mengubah kata menjadi kata dasarnya dengan menghilangkan imbuhan-imbuhan pada kata dalam dokumen atau mengubah kata kerja menjadi kata benda . Stem (akar kata) adalah kata inti setelah imbuhan dihilangkan (awalan dan akhiran)

d. Tahap Tagging :Tahap ini merupakan tahapan untuk mencari bentuk awal atau root dari tiap-tiap kata lampau atau kata dari hasil stemming

Pembobotan Kata

Pembobotan kata adalah sebuah proses dari pemberian nilai bobot berdasarkan term indeks. Ada beberapa pembobotan kata yang bisa digunakan, namun dipenelitian ini peneliti menggunakan pembobotan kata TF-IDF. TF IDF merupakan hasil perkalian dari Term Frequency atau jumlah kemunculan kata pada tiap dokumen serta Inverse Document Frequency atau kemunculan sebuah term dalam dokumen yang paling sedikit. Rumus dari TF IDF ditunjukkan oleh Rumus 1 yaitu :

$$W_{t,d} = W_{tft,d} x\, idf_t \quad (1)$$

a. *Term Frequency* (TF)

Term frequency adalah seberapa sering kemunculan kata pada satu dokumen[10]. Rumus TF yang ditunjukkan oleh Rumus 2 adalah sebagai berikut :

$$W_{tf_{i,d}} = \begin{cases} 1 + log_{10}\, tf_{t,d}, if\ tf_{t,d} > 0 \\ 0 \end{cases} \quad (2)$$

Keterangan :

$Tf_{i,d}$ adalah jumlah kemuculan term t di dokumen d.

b. *Document Frequency* (DF)

Documenn requency merupakan kata yang sering muncul di dokumen[10]. Contoh : dan, di, atau, bisa

c. Invers Document Frequency (IDF)

Invers document frequency adalah kemungkinan munculnya term diseluruh dokumen[10]. Maka term yang jarang sekali muncul, nilai bobot IDF semakin besar. Berikut adalah rumus dari IDF yang ditunjukkan oleh Rumus 3:

$$idf_t = \log 10 + \frac{N}{df_{(t)}} \quad (3)$$

d. Term Frequency – Invers Document Frequency (TF-IDF)

Metode TF-IDF menggabungkan dua konsep untuk menghitung bobot, yaitu frekuensi kemunculan suatu kata dalam dokumen tertentu dan frekuensi kebalikan dari dokumen yang mengandung kata tersebut. [5]. Frekuensi kemunculan kata di dalam dokumen yang diberikan menunjukkan seberapa penting kata itu di dalam dokumen tersebut. Frekuensi dokumen yang mengandung kata tersebut menunjukkan seberapa umum kata tersebut.

Untuk rumus pembobotan text dengan metode TF-IDF ditunjukkan oleh Rumus 4.

$$W_{i,j} = tf_{i,j} .\log\left(\frac{N}{df_i}\right) \quad (4)$$

Naïve Bayes Classifier

Naïve Bayes Classifier merupakan metode yang berfungsi untuk mrnghitung nilai probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari data set[11]. Keuntungan menggunakan *Naïve Bayes* adalah metode ini hanya membutuhkan dua data yaitu data training (training set) dan data Testing (Testing set) untuk menguji suatu data yang ingin diperoleh yang ditunjukkan oleh Rumus 5.

$$P(Cj|Wi) = \frac{P(Cj)\ x\ P(Wi|Cj)}{P(Wi)} \quad (5)$$

Menghitung jumlah dokumen pada kategori tertentu digambarkan pada persamaan Rumus 6 berikut:

$$P(C_j) = N(C_j) / N \quad (6)$$

Multinomial Model merupakan model probabilitas yang penelti gunakan. Berikut merupakan persamaan *Multinomial Model* yang ditunjuukan oleh Rumus 7.

$$P(w|c) = \frac{Count\,(w,c)+1}{Count(c)+|V|} \quad (7)$$

Metode Pengujian

Adapun tiga tahapan metode pengujian yang dipakai dalam penelitian ini adalah sebagai berikut :

a. Confusion Matrix

Confusion matrix atau *error matrix* adalah sebuah metode perhitungan akurasi terhadap sebuah sistem pada konsep data mining. Terdapat 4 istilah di dalam *Confusion matrix* yaitu, True Positif (TP), True Negatif (TN), False Positif (FP), dan False Negatif (FN).

b. *Precision and Recall*

Precision adalah tingkat presisi antara informasi yang diminta dan respons yang diberikan oleh sistem. Sedangkan *Recall* adalah tingkat keberhasilan sistem dalam mengambil informasi.

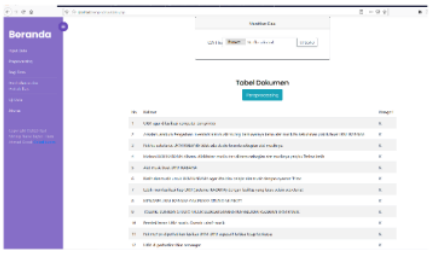
c. *Cross Validation*

Cross Validation adalah metode statistik untuk mengukur

kinerja model algoritma dimana data dipisahkan menjadi 2 subset, yaitu data latih dan data uji. Pada penelitian ini penulis menggunakan 10 k-fold cross validation.

HASIL DAN PEMBAHASAN

Hasil dari penelitian ini adalah sebuah sistem klasifikasi keluhan mahasiswa yang dapat secara otomatis mengklasifikasi keluhan-keluhan mahasiswa ke unit kerja tujuan keluhan. Adapun tampilan dari sistem yang dihasilkan dapat dilihat pada gambar 2 berikut:

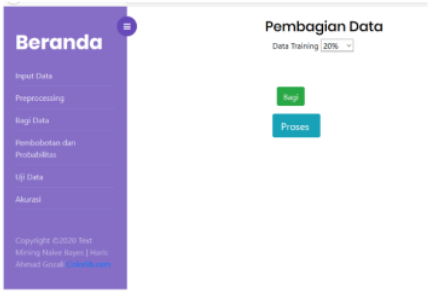


Gambar 2. *Tampilan data keluhan mahasiswa*

Sebelum sistem ini dapat mengklasifikasi data keluhan mahasiswa, harus melalui tahap *preprocessing* terlebih dahulu melalui menu *preprocessing* yang ditunjukkan oleh gambar 3.

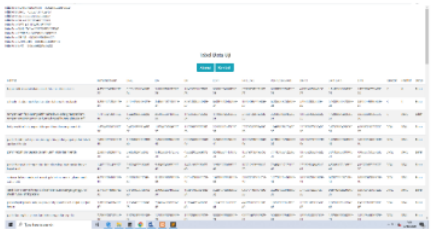
Gambar 3. *Tampilan Halaman Preprocessing*

Setelah melalui tahap proses *preprocessing* data dibagi menjadi 2 yaitu data *training* dan data *testing* seperti gambar 4 berikut:



Gambar 4. *Pembagian data training dan testing*

Pada pembagian data ini, data keluhan akan diambil secara acak. Pada penelitian ini menggunakan data training 70% dari *dataset*. Kemudian akan dicoba klasifikasi data testing terhadap data training yang sudah melalui pembobotan dan perhitungan probabilitas seperti gambar 5 berikut:



Gambar 5. *Hasil klasifikasi data testing*

Dari gambar 5 terlihat bahwa sistem berhasil melakukan klasifikasi keluhan mahasiswa, ada yang dapat terklasifikasi dengan benar sesuai dengan klasifikasi manual, ada yang terklasifikasi salah. Untuk mengukur nilai akurasi dilakukan dengan menggunakan metode cross validation, recall and precission dan confusion matrix, pada penelitian ini hasil ketiga metode tersebut dapat dilihat pada gambar 6,7 dan 8 sebagai berikut:

Gambar 6. *Tampilan uji coba dengan metode confusion matrix*

Gambar 7. *Tampilan uji coba dengan precission dan recall*

Gambar 8. *Tampilan uji coba dengan metode cross*

validation

Terlihat bahwa pada hasil uji coba, sistem yang dihasilkan memiliki tingkat akurasi rata-rata 82%.

KESIMPULAN

Metode *naïve bayes* dan library sastrawi dapat melakukan klasifikasi keluhan mahasiswa dengan baik dan dapat memperbaiki nilai akurasi pada penelitian sebelumnya yang memiliki nilai akurasi rata-rata 79% menjadi 82% menggunakan *naïve bayes* dan library sastrawi pada proses *preprocessing*-nya.

REFERENSI

[1] Torayev, P. C. M. M. Magusin, C. P. Grey, C. Merlet, and A. A. Franco, " Text mining assisted review of the literature on Li-O 2 batteries ," *J. Phys. Mater.*, vol. 2, no. 4, p. 044004, 2019.

[2] S. Gusriani, K. D. K. Wardhani, and M. I. Zul, "Analisis Sentimen Terhadap Toko Online di Sosial Media Menggunakan Metode Klasifikasi Naïve Bayes (Studi Kasus: Facebook Page BerryBenka)," *4th Appl. Bus. Eng. Conf.*, vol. 1, no. 1, pp. 1–7, 2016.

[3] Spr. Rani, B. Ramesh, and M. Anusha, "Evaluation of stemming techniques for text classification," *J. Comput. ...*, vol. 43, no. 3, pp. 165–171, 2015.

[4] M. A. Rosid, G. Gunawan, and E. Pramana, "Centroid Based Classifier With TF – IDF – ICF for Classfication of Student's Complaint at Appliation E-Complaint in Muhammadiyah University of Sidoarjo," vol. 1, no. 1, 2015.

[5] R. T. Wahyuni, D. Prastiyanto, and E. Supraptono, "Penerapan Algoritma Cosine Similarity dan Pembobotan TF-IDF pada Sistem Klasifikasi Dokumen Skripsi," *J. Tek. Elektro*, vol. 9, no. 1, pp. 18–23, 2017.

[6] A. Rachmat C and Y. Lukito, "Klasifikasi Sentimen Komentar Politik dari Facebook Page Menggunakan Naive Bayes," *J. Inform. dan Sist. Inf. Univ. Ciputra*, vol. 02, no. 02, pp. 26–34, 2016.

[7] L. Marlina, M. lim, and A. P. Utama Siahaan, "Data Mining Classification Comparison (Naïve Bayes and C4.5 Algorithms)," *Int. J. Eng. Trends Technol.*, vol. 38, no. 7, pp. 380–383, 2016.

[8] J. T. Informasi *et al.*, "PENGARUH TEXT PREPROCESSING DAN KOMBINASINYA," vol. 15, pp. 1–11, 2019.

[9] S. Vijayarani, M. J. Ilamathi, M. Nithya, A. Professor, and M. P. Research Scholar, "Preprocessing Techniques for Text Mining -An Overview," vol. 5, no. 1, pp. 7–16.

[10] W. E. Nurjanah, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan Metode K-Nearest Neighbor dan Analisis Sentimen Terhadap Tayangan Televisi Berdasarkan Opini Masyarakat pada Media Sosial Twitter menggunakan M," no. October, 2017.

[11] S. S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms," *Int. J. Mod. Trends Eng. Res.*, vol. 4, no. 7, pp. 58–63, 2017.

Conflict of Interest Statement:
The author declares that the research was conducted in the absence of any commercial or financialrelation- ships that could be construed as a potential conflict of interest.

Article History:

Received: 2020-01-23 | Accepted: 2020-03-30 | Published: 2020-04-29

unplug Classification of Student Complaints with Naive Bayes and Literary Methods

ORIGINALITY REPORT

16%	16%	14%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journal.uc.ac.id Internet Source	1%
2	pubs.rsc.org Internet Source	1%
3	Submitted to Imperial College of Science, Technology and Medicine Student Paper	1%
4	ejournal.unib.ac.id Internet Source	1%
5	Submitted to Universitas Dian Nuswantoro Student Paper	1%
6	download.garuda.kemdikbud.go.id Internet Source	1%
7	jurnal.stkippgritulungagung.ac.id Internet Source	1%
8	I Putu Dedy Wira Darmawan, Gede Aditra Pradnyana, Ida Bagus Nyoman Pascima. "Optimasi Parameter Support Vector Machine	1%

Dengan Algoritma Genetika Untuk Analisis Sentimen Pada Media Sosial Instagram", SINTECH (Science and Information Technology) Journal, 2023
Publication

9	journal.untar.ac.id Internet Source	1 %
10	openlibrarypublications.telkomuniversity.ac.id Internet Source	1 %
11	Submitted to Universitas Riau Student Paper	1 %
12	elibrary.unikom.ac.id Internet Source	1 %
13	Surohman Surohman, Sopian Aji, Rousyati Rousyati, Fanny Fatma Wati. "Analisa Sentimen Terhadap Review Fintech Dengan Metode Naive Bayes Classifier Dan K- Nearest Neighbor", EVOLUSI : Jurnal Sains dan Manajemen, 2020 Publication	1 %
14	ejournal.bsi.ac.id Internet Source	1 %
15	Agatha Deolika, Kusrini Kusrini, Emha Taufiq Luthfi. "ANALISIS PEMBOBOTAN KATA PADA KLASIFIKASI TEXT MINING", JURNAL TEKNOLOGI INFORMASI, 2019 Publication	1 %

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		