

IMPLEMENTASI TEXT MINING MENGGUNAKAN *NAIVE BAYES* UNTUK PENENTUAN KATEGORI TUGAS AKHIR MAHASISWA BERDASARKAN ABSTRAKSINYA

Agus Hermanto

Teknik Informatika, Fakultas Teknik, Universitas 17 Agustus 1945 Surabaya
Email: hermantto_if@untag-sby.ac.id

ABSTRAK

Dengan semakin meningkatnya penggunaan aplikasi sistem informasi di berbagai bidang, turut berdampak pada kebutuhan untuk peningkatan kecepatan pemrosesan data. Pemrosesan data yang menggunakan proses semi manual, mempunyai beberapa kendala, diantaranya : waktu proses lebih lama dan besaran data yang diproses menjadi relatif sedikit. Oleh karena itu dalam penelitian ini dikembangkan penggunaan *naive bayes* untuk membantu bagian koordinator tugas akhir dalam melakukan pengelompokan proposal tugas akhir. Metode *naive bayes* yang akan diimplementasi ke dalam sistem informasi proposal tugas akhir dapat memberikan sebuah solusi baru untuk menentukan kategori proposal tugas akhir berdasarkan abstraksi yang dibuat mahasiswa. Dalam hasil uji coba metode ini, dapat disimpulkan cukup berhasil dan secara garis besar dapat dijadikan sebagai perangkat bantu dalam melakukan klasifikasi dokumen tugas akhir. Tingkat akurasi berdasarkan pengujian untuk kategori hardware dan networking mencapai 86%, kategori sistem informasi tingkat akurasi mencapai 80% dan kategori sistem informasi akuntansi mencapai 89%. Secara keseluruhan, berdasarkan jumlah dataset yang diujikan dan tingkat keberhasilan yang dicapai, maka sistem ini mempunyai tingkat akurasi 87%.

Kata Kunci : *naive bayes classifier*, machine learning, supervised learning, information retrieval, text mining

1. PENDAHULUAN

Perkembangan sistem informasi dalam beberapa dekade terakhir menunjukkan hasil menggembirakan dengan hasil berbagai informasi yang semakin variatif dan kompleks.

Berbagai metode dibenamkan kedalam *engine* dari sistem informasi untuk meningkatkan utilitas dari sistem informasi itu sendiri.

Salah satu permasalahan yang ditangani oleh sistem informasi adalah *information retrieval* atau penemuan kembali informasi menggunakan *text mining*. Pemanfaatan text mining ini

sangat bervariasi, diantaranya adalah sebagai *classifier*, yaitu perangkat bantu untuk mengklasifikasi data.

Penggunaan *classifier* dalam aplikasi dapat meningkatkan utilitas sistem operasi berupa kecepatan proses pencarian karena data telah dikempokkan berdasarkan kategori tertentu secara signifikan, dengan menggunakan proses *machine learning* dengan cara *supervised* ataupun *unsupervised*.

Obyek penelitian adalah kumpulan abstraksi dari tugas akhir mahasiswa yang sudah lulus dan akan diunggah ke sistem repository tugas

akhir. Selama ini Politeknik XYZ belum mempunyai sistem repository digital yang dipergunakan dalam menyimpan tugas akhir mahasiswa, sehingga kumpulan tugas akhir berupa kepingan *compact disc* yang semakin banyak menumpuk di perpustakaan, dari tahun ke tahun. Tentunya, hal ini tidak dapat dimanfaatkan oleh para mahasiswa yang masih menempuh perkuliahan sebagai bahan referensi untuk dijadikan pertimbangan memilih topik tugas akhir.

Berdasarkan pertimbangan tersebut, maka Politeknik XYZ akhirnya memutuskan untuk membuat sebuah sistem repository untuk tugas akhir mahasiswa.

Sistem repository yang akan dikembangkan harus mampu memproses data dengan cepat dikarenakan banyaknya karya tugas akhir mahasiswa yang akan diunggah. Oleh karena itu perlu untuk membangun sistem yang dilengkapi dengan kemampuan machine learning sebagai solusinya.

Penelitian kali ini berfokus pada perancangan dan simulasi perangkat aplikasi untuk pengklasifikasian laporan tugas akhir mahasiswa berdasarkan artikel abstraksinya, sedangkan metode yang digunakan adalah supervised learning dengan *naive bayes classifier*, sebagai fitur utama dari sistem.

2. LANDASAN TEORI

Pembelajaran mesin adalah jenis kecerdasan buatan (AI) yang menyediakan komputer dengan kemampuan untuk belajar tanpa diprogram secara eksplisit (Norvig & Russel, 2003). Pembelajaran mesin berfokus pada pengembangan program komputer yang dapat berubah ketika memperoleh data baru untuk diproses.

Bidang pembelajaran mesin berkembang dari studi pengenalan pola dan teori belajar komputasi dalam

kecerdasan buatan, pembelajaran mesin mengeksplorasi studi dan pembuatan algoritma yang dapat belajar dari data dan membuat prediksi pada data tersebut (Mackay, 2003).

Pembelajaran mesin yang digunakan dalam berbagai tugas komputasi untuk perancangan dan pemrograman algoritma eksplisit yang tidak memungkinkan, misalnya aplikasi penyaringan spam, deteksi penyusup jaringan atau optical character recognition (OCR), mesin pencari dan visi komputer.

Proses pembelajaran pada mesin terkait erat dengan (dan sering tumpang tindih dengan) statistik komputasi, yang juga berfokus pada prediksi keputusan melalui penggunaan komputer. Bidang ini memiliki ikatan yang kuat untuk optimasi matematika, yang memberikan metode, teori dan aplikasi domain ke field tertentu. Pembelajaran mesin kadang-kadang digabungkan dengan data mining, dimana subfield terakhir lebih memfokuskan pada analisis data eksplorasi dan dikenal sebagai *unsupervised learning*, yang digunakan untuk proses belajar dan membangun profil perilaku dasar untuk berbagai entitas, sehingga dapat digunakan untuk menemukan anomali yang berarti (Norvig & Russel, 2003).

Proses pembelajaran mesin dapat diklasifikasikan ke dalam tiga kategori besar (Norvig & Russel, 2003), yaitu :

- a. **Supervised learning:** Komputer disajikan dengan contoh input dan output yang diinginkan, diberikan oleh seorang "pakar", dan tujuannya adalah untuk mempelajari aturan umum yang memetakan input ke output.
- b. **Unsupervised learning:** Tidak ada label yang diberikan kepada algoritma pembelajaran, sehingga komputer melakukan sendiri untuk

menemukan struktur dalam input. Belajar tanpa pengawasan dapat menjadi tujuan dalam dirinya sendiri (menemukan pola tersembunyi dalam data) atau menemukan jalan menuju akhir (fitur learning).

- c. **Reinforcement learning:** Sebuah program komputer berinteraksi dengan lingkungan yang dinamis. Program tersebut harus melakukan suatu tujuan tertentu (seperti mengendarai kendaraan atau bermain game melawan lawan).

Dalam machine learning, istilah *naive bayes classifier* adalah keluarga dari pengklasifikasi probabilistik sederhana berdasarkan menerapkan teorema Bayes 'dengan asumsi kuat (naif), yang berupa asumsi kemandirian diantara fiturnya.

Metode naive bayes classifier sangat scalable (Nigam, McCallum, Thrun, & Mitchell, 1998), hanya membutuhkan sejumlah parameter linier dalam jumlah variabel (fitur atau prediktor) dalam masalah belajar. pelatihan kemungkinan maksimum bisa dilakukan dengan mengevaluasi ekspresi bentuk tertutup, yang yang membutuhkan waktu linier, bukan dengan pendekatan berulang mahal seperti yang digunakan untuk jenis lain dari pengklasifikasi.

Dalam statistik dan ilmu komputer, model Naive Bayes yang dikenal dengan berbagai nama, diantaranya simple bayes dan independence bayes, yang mereferensikan terhadap penggunaan teorema bayes dalam aturan keputusan classifier, tapi naive bayes tidak (selalu) merujuk pada metode Bayesian (Hand & Yu, 2001).

Naive Bayes adalah teknik yang sederhana untuk membangun model pengklasifikasi yang menetapkan kelas pelabelan untuk contoh masalah, yang direpresentasikan sebagai vektor dari

nilai-nilai fitur, di mana label kelas itu diambil dari beberapa himpunan yang berhingga. Misalnya, suatu buah dapat dianggap jeruk jika berwarna kuning, oval, dan dengan diameter sekitar 7 cm. Sebuah naive bayes classifier menganggap masing-masing fitur ini untuk berkontribusi secara independen untuk probabilitas bahwa buah ini adalah jeruk, terlepas dari korelasi yang mungkin terdapat diantara fitur warna, kebulatan, dan diameter.

Untuk beberapa jenis model probabilitas, naive bayes classifier dapat dilatih dengan sangat efisien dalam pengaturan supervised learning. Dalam banyak aplikasi praktis, estimasi parameter untuk model naive bayes menggunakan metode probabilitas maksimum, dengan kata lain, kita dapat bekerja dengan model naive bayes tanpa mengetahui konsep probabilitas Bayesian atau menggunakan metode Bayesian.

Hasil perbandingan yang komprehensif dengan algoritma klasifikasi lainnya, menunjukkan bahwa naive bayes lebih unggul dibandingkan berbagai pendekatan lainnya, seperti boosted trees or random forests (Caruana & Niculescu-mizil, 2006). Kelebihan dari naive bayes classifier adalah bahwa metode ini hanya membutuhkan sejumlah kecil data pelatihan untuk memperkirakan parameter yang diperlukan dalam klasifikasi.

Secara abstrak, naive bayes classifier adalah model probabilitas bersyarat (Murty & Devi, 2011): diberikan contoh masalah yang harus diklasifikasikan, diwakili oleh vektor $x = \{x_1, \dots, x_n\}$ yang mewakili beberapa fitur n (variabel independen), akan ditetapkan untuk nilai probabilitasnya, yaitu :

$$p(C_k | x_1, \dots, x_n) \quad (\text{pers. 1})$$
 untuk setiap kemungkinan hasil k atau kelas C_k .

Permasalahan dengan formula di atas adalah bahwa jika sejumlah fitur n besar atau jika fitur dapat mengambil sejumlah besar nilai n , maka model tersebut adalah tidak layak berdasarkan pada tabel probabilitas. Oleh karena itu perlu merumuskan model untuk membuatnya lebih sederhana. Dengan menggunakan teorema Bayes, probabilitas bersyarat dapat diuraikan sebagai berikut :

$$P = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \quad (\text{pers. 2})$$

atau dalam bentuk notasi matematika :

$$P(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (\text{pers. 3})$$

Dalam prakteknya, hanya menarik di pembilang dari fraksi itu, karena penyebut tidak tergantung pada C dan nilai-nilai dari fitur F_i yang diberikan, sehingga penyebut dikatakan efektif berupa nilai konstan dan pembilang setara dengan model probabilitas gabungan.

Contoh implementasi naive bayes classifier dapat dijelaskan berdasarkan beberapa penelitian yang telah ada sebelumnya, antara lain:

- a. Penggunaan naive bayes sebagai filter spam pada email (Chandra, Indrawan, & Sukajaya, 2016). Dalam penelitian ini, metode naive bayes dikombinasikan dengan POS Tagger. Hasil penelitian menunjukkan nilai akurasi yang dicapai berkisar antara 78.72% sampai dengan 84.30%, sehingga masih tergolong dalam tingkat *good classification* (Gorunescu, 2011).
- b. Penggunaan naive bayes untuk pengklasifikasian trending topic twitter (Agustina, Matulatan, & Bettiza, 2014). Dalam penelitian, dilakukan penggolongan trend ke dalam kategori tertentu tersebut menerapkan metode Naïve Bayes. Penggolongan trend dibagi menjadi 5 (lima) kategori. Hasil yang

diperoleh dengan menggunakan data uji menunjukkan kategori Religi 16,67%, Olahraga 36,7%, Berita 6,7%, Televisi & Film 6,7% dan Musik 33,3%. Berdasarkan hasil tersebut, pengklasifikasian trending topic dapat mempunyai nilai manfaat tambahan, misalnya menentukan segmen iklan sehingga lebih tepat sasaran.

- c. Penelitian berikut ini menggunakan naive bayes dan algoritma C4.5 dalam *decision tree* untuk memprediksi perilaku klien (Karim & Rahman, 2013). Dalam penelitian ini telah dilakukan pengujian dua teknik data mining: yang Naïve Bayes dan algoritma pohon keputusan menggunakan C4.5. Tujuannya adalah untuk memprediksi apakah klien akan berlangganan deposito berjangka. Penelitian ini juga membuat studi perbandingan kinerja dari kedua algoritma. Data Publik UCI yang tersedia digunakan untuk melatih dan menguji kinerja algoritma. Penelitian ini juga berhasil melakukan ekstraksi pengetahuan ditindaklanjuti dari pohon keputusan yang berfokus pada pengambilan keputusan yang menarik dan penting di daerah bisnis.

3. METODOLOGI PENELITIAN

Politeknik XYZ mempunyai 3 (tiga) program studi dengan jenjang Diploma III, yaitu teknik komputer, manajemen informatika dan komputerisasi akuntansi. Keseluruhan program studi tersebut mempunyai kebebasan dalam hal kategori topik tugas akhirnya, sehingga tidak akan dikelompokkan berdasarkan program studi, melainkan berdasarkan topik-topik yang berkaitan dengan bidang teknologi informasi dan keuangan atau akuntansi. Adapun kategori yang digu-

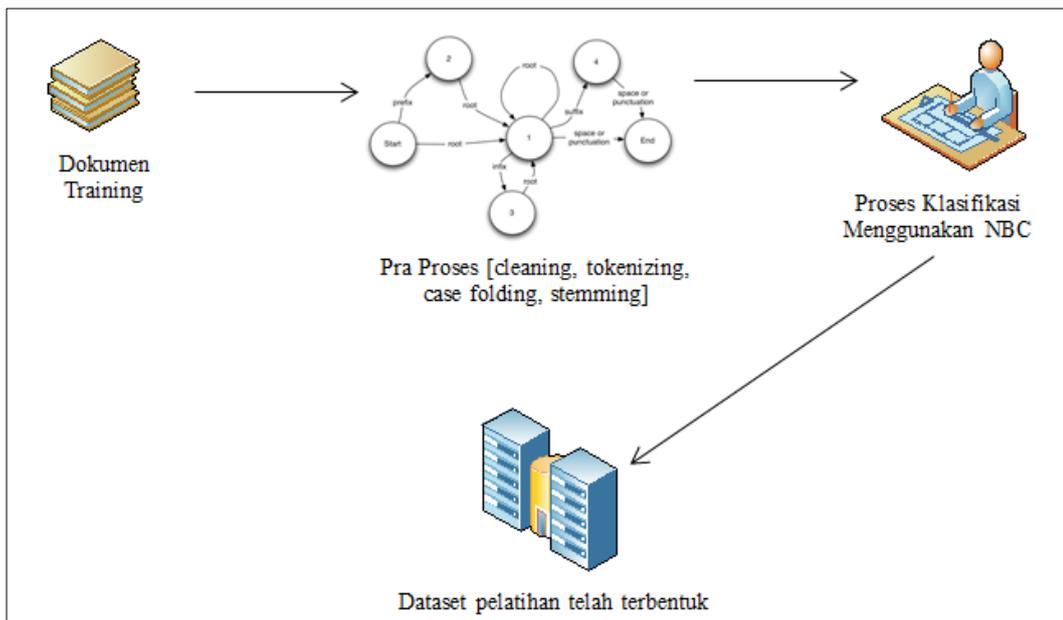
nakan dalam penelitian ini adalah Hardware dan Networking, Sistem Informasi dan Sistem Informasi Akuntansi. Pengambilan ketiga kategori ini didasarkan pada pengambilan sampel acak beberapa file tugas akhir dan diklasifikasikan manual. Berikut adalah penjelasan mengenai kategori-kategori tersebut :

a. Kategori Hardware dan Networking. Kategori ini berhubungan

pemrograman web dan aplikasi mobile.

c. Kategori sistem informasi akuntansi. Kategori ini berhubungan dengan perpajakan, point of sales, keuangan dan akuntansi.

Langkah selanjutnya dalam melakukan pengklasifikasian proposal tugas akhir ini, dibutuhkan bahan penelitian yang berupa file yang berisi abstraksi, yang akan dibagi menjadi



Gambar 1. Proses Pembentukan Dataset Dari Dokumen Pengujian

dengan jaringan komputer, keamanan jaringan dan data, topologi jaringan, pemulatan alat, penggunaan sensor, dan mikrokontroler.

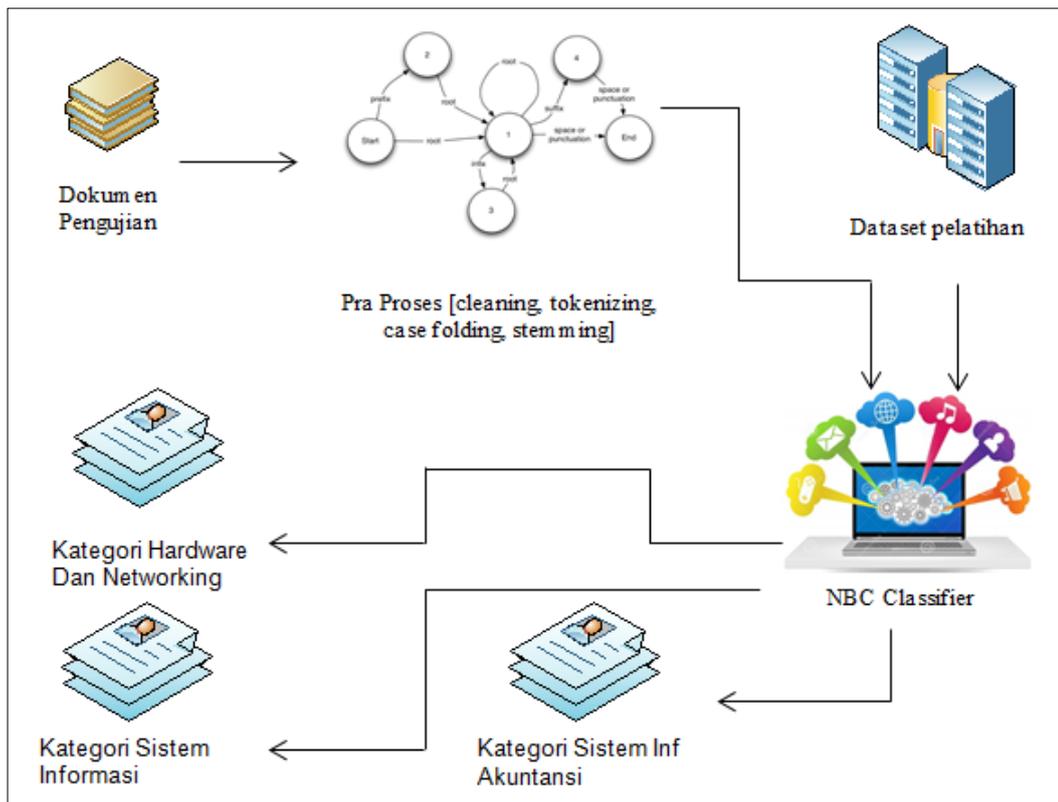
b. Kategori Sistem Informasi. Kategori ini berhubungan dengan basis data, database, pembelajaran interaktif dan multimedia, termasuk

data pelatihan dan data pengujian.

Adapun tahapan dalam membuat *classifier* dapat dilihat pada gambar 1, dengan diawali dengan pra proses. Tahap ini bertujuan untuk menyiapkan data sebelum digunakan pada proses klasifikasi, seperti yang terlihat pada gambar 1.

Pemrosesan data pada praproses meliputi pembersihan tanda baca dan simbol selain alphabet, kemudian dila-

bentuk dasar yang sama. Ketika proses stemming ini selesai dilakukan, maka data tersebut telah siap untuk digu-



Gambar 2. Proses Pengklasifikasian Menggunakan Data Pengujian

kukan proses pemotongan file abstrak menjadi beberapa bagian, yang diistilahkan dengan token. Setiap token tersebut dilakukan perubahan menjadi *lower case* semua. Proses ini dinamakan dengan *casefolding*. Tujuan dari proses tersebut adalah mempermudah pencarian, karena teks cenderung terdapat huruf besar dan kecil. Langkah akhir dari tahap praproses ini adalah menghilangkan *stop word*, karena kata-kata yang termasuk dalam stopword tidak terlalu bermakna pada isi dokumen.

Tahapan berikutnya adalah melakukan proses stemming, yaitu proses menghapus imbuhan berupa *prefixes* (awalan), *infix* (sisipan) dan *suffixes* (akhiran), sehingga didapatkan kata dasarnya. Hal ini bertujuan memperkecil jumlah variasi yang mempunyai

nakan dalam proses pelatihan dan klasifikasi.

Proses pelatihan dan klasifikasi menggunakan naive bayes classifier, pada dasarnya menggunakan prinsip probabilitas sederhana yang diambil dari teori Bayes (statistik Bayesian), tentunya hal ini berdasarkan asumsi naif atau independen yang kuat.

Pada tahap pelatihan ini dilakukan proses pengolahan menggunakan beberapa dokumen contoh, sedangkan tahap pengujian dilakukan proses klasifikasi dokumen yang baru supaya dapat diperoleh informasi kategori dari dokumen tersebut.

Penggunaan naive bayes classifier menggunakan asumsi bahwa antara fitur yang satu dengan fitur yang lain tidak berhubungan.

Hasil proses pelatihan yang berupa dataset ini nantinya akan digunakan sebagai dasar pemrosesan ketika menggunakan dokumen baru, seperti yang terlihat pada gambar 2.

Berikut ini adalah rancangan *algorithmna naive bayes classifier* yang digunakan dalam penelitian ini:

- a. Inisialisasi bobot nilai probabilitas kategori dokumen menjadi $1/3$.
- b. Alokasikan besaran memori sesuai dengan jumlah hasil penghitungan buffer dokumen yang diproses.
- c. Inisialisasi nilai probabilitas dokumen menjadi maksimum.
- d. Lakukan proses query pada dataset Model dengan memilih data yang berada pada kolom kategori 1.
- e. Lakukan iterasi mulai awal token sampai akhir untuk melakukan proses penghitungan bobot dokumen sesuai token aktif yang diproses, untuk dokumen kategori 1.
- f. Lakukan query pada dataset model untuk menampilkan data-data yang berada pada kolom kategori 2.
- g. Lakukan iterasi mulai awal token sampai akhir untuk melakukan proses penghitungan bobot dokumen sesuai token aktif yang diproses, untuk dokumen kategori 2.
- h. Lakukan query pada dataset model untuk menampilkan data-data yang berada pada kolom kategori 3;
- i. Tentukan kategori dokumen yang sedang diproses dengan memeriksa nilai probabilitasnya yang terbesar.
- j. Tampilkan hasil penentuan kategori dokumen.

Cara kerja *algorithmna* diatas pada dasarnya adalah bayes menggunakan prinsip probabilitas dalam menentukan suatu keputusan, dengan melihat nilai kemungkinan yang terbesar. Dalam permasalahan ini, terdapat tiga kategori tugas akhir, yaitu hardware dan networking, sistem informasi dan sistem infotmasi akuntansi. Mengingat tetap ada kemungkinan munculnya nilai nol dalam perhitungan probabilitas bayes, maka secara prinsip tetap ditambahkan satu kategori, yaitu tidak diketahui, jika hasil perhitungan bernilai nol (walaupun kondisi ini sangat jarang terjadi).

4. HASIL DAN PEMBAHASAN

Sebelum melakukan proses pengujian, terlebih dahulu menyiapkan dokumen pelatihan, seperti yang tampak pada tabel 1. Dokumen pelatihan tersebut telah diklasifikasikan secara manual terlebih dahulu sebagai metode untuk memberikan pembelajaran awal pada *naive bayes engine*-nya.

Kemudian dataset pelatihan tersebut di-entry ke sistem dan disimpan. Setelah itu, data set tersebut dilakukan proses pelatihan oleh sistem. Setelah proses pelatihan telah selesai dilakukan, sistem dapat digunakan untuk proses pengujian klasifikasi tugas akhir.

Tahap berikutnya adalah menyiapkan dokumen pengujian, yang belum diketahui kategorinya oleh sistem. Dokumen yang telah diuji proses klasifikasinya akan diuji coba silang pada proses pengujian untuk kategori yang lain, sehingga dapat diketahui kemampuan sistem dan tingkat akurasinya dengan baik.

Tabel 1, Daftar Dokumen Pelatihan

No	Kode_dok	Kategori
1	DU 064/09	Hardware And Networking
2	DU 031/09	Hardware And Networking
3	DU 111/010	Hardware And Networking
4	DU 002/08	Hardware And Networking
5	DU 045/08	Hardware And Networking
6	DU 067/09	Hardware And Networking
7	DU 155/010	Hardware And Networking
8	DU 011/09	Sistem Informasi
9	DU 015/011	Sistem Informasi
10	DU 017/08	Sistem Informasi
11	DU 056/09	Sistem Informasi
12	DU 090/010	Hardware And Networking
13	DU 205/011	Sistem Informasi Akuntansi
14	DU 166/011	Sistem Informasi Akuntansi
15	DU 183/011	Sistem Informasi Akuntansi
16	DU 094/09	Sistem Informasi Akuntansi
17	DU 016/010	Sistem Informasi
18	DU 129/011	Sistem Informasi
19	DU 055/09	Sistem Informasi
20	DU 040/09	Sistem Informasi
21	DU 064/09	Hardware And Networking
22	DU 031/09	Hardware And Networking
23	DU 111/010	Sistem Informasi Akuntansi
24	DU 002/08	Sistem Informasi Akuntansi
25	DU 045/08	Sistem Informasi Akuntansi
26	DU 067/09	Sistem Informasi Akuntansi
27	DU 155/010	Sistem Informasi Akuntansi
28	DU 011/09	Sistem Informasi
29	DU 015/011	Sistem Informasi
30	DU 017/08	Sistem Informasi Akuntansi

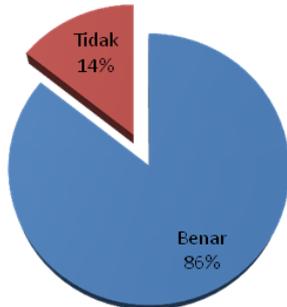
Tabel 2. Daftar Dokumen Pengujian

No	Kategori	# Dokumen
1	Sistem Informasi	35
2	Hardware And Networking	35
3	Sistem Informasi Akuntansi	35
	Jumlah Total	105

Hasil pengujian untuk dokumen pengujian tiap kategori dapat dijelaskan sebagai berikut :

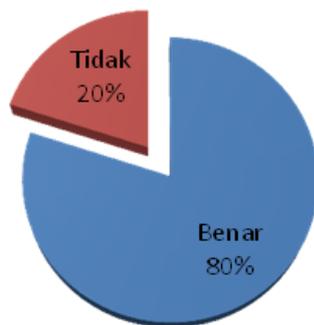
- a. Dokumen pengujian untuk kategori Hardware dan Networking. Sistem ini memproses 35 dokumen yang asalnya belum terkategori. Hasil pengujian menggunakan *naive ba-*
- yes* menunjukkan 30 dokumen berhasil diproses klasifikasinya dengan benar, tampak pada gambar 3.
- b. Dokumen pengujian untuk kategori Sistem Informasi. Sistem ini memproses 35 dokumen yang asalnya belum terkategori. Hasil pengujian menggunakan *naive bayes* menun-

jujukan 28 dokumen dalam prose pengujian ini berhasil diproses klasifikasinya dengan benar, terlihat pada gambar 4.



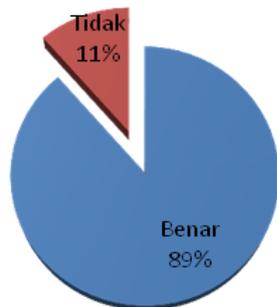
Gambar 3. Dokumen Kategori Hardware & Networking

- c. Dokumen pengujian untuk kategori Sistem Informasi Akuntansi. Sistem ini memproses 35 dokumen yang asalnya belum terkategori. Hasil dengan pengujian menggunakan *naive bayes* menunjukkan 33 dokumen berhasil diproses klasifi-

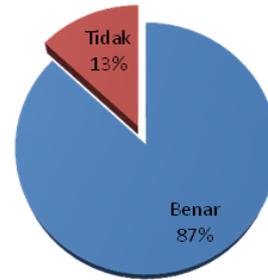


Gambar 4. Pengujian Dokumen Kategori Sistem Informasi

klasifikasinya dengan benar, seperti yang terlihat pada gambar 6.



Gambar 5. Pengujian Dokumen Kategori Sistem Informasi Akuntansi



Gambar 6. Tingkat Keberhasilan Pengujian Seluruh Dokumen

Berikut ini adalah daftar status hasil pengujian, seperti yang terlihat pada tabel 3 berikut ini :

Tabel 3. Hasil Pengujian

Kategori	# Status	
	Benar	Salah
Sistem Informasi	28	7
Hardware And Networking	30	5
Sistem Informasi Akuntansi	33	2
Jumlah	91	14

Berdasarkan tabel diatas, dapat diketahui bahwa keberhasilan paling tinggi pada kategori Sistem Informasi Akuntansi, kemudian Hardware dan Networking, dan Sistem Informasi.

Kesalahan yang terjadi pada proses pengujian klasifikasi dikarenakan adanya beberapa kata kunci yang mirip antar kategori, baik untuk kelompok hardware and networking, sistem informasi, dan sistem informasi akuntansi.

5. PENUTUP

Berdasarkan hasil pengujian dapat disimpulkan bahwa pengelompokan tugas akhir dengan menggunakan *naive bayes* ini berhasil dengan cukup baik dengan tingkat akurasi yang berhasil dicapai dari 105 kali pengujian, adalah 91 berhasil dan 14 kali gagal atau sekitar 87 %, sehingga dapat dikatakan penggunaan *naive bayes* sudah cukup tepat dalam melakukan pengelompokan tugas akhir berdasarkan abstraksinya.

Pengelompokkan tugas akhir ini masih dapat dikembangkan seiring dengan berkembangnya spesifikasi kebutuhan pengguna sistem yang harus dipenuhi dalam mencapai hasil dan kinerja sistem yang lebih baik, dengan mengkombinasikan metode lain untuk meningkatkan keberhasilan dalam melakukan klasifikasi.

DAFTAR PUSTAKA

1. Agustina, P. A., Matulatan, T., & Bettiza, M. (2014). The Classification Of The Trending Topic Of Twitter's With Naïve Bayes Method. *Jurnal UMRAH (Universitas Maritim Raja Ali Haji)*.
2. Caruana, R., & Niculescu-mizil, A. (2006). An Empirical Comparison of Supervised Learning Algorithms. *In Proc. 23 rd Intl. Conf. Machine learning (ICML '06)* (hal. 161--168). Pittsburgh, Pennsylvania, USA: ACM New York, NY, USA.
3. Chandra, W. N., Indrawan, G., & Sukajaya, I. (2016). Spam Filtering Dengan Metode Pos Tagger Dan Klasifikasi Naïve. *Jurnal Ilmiah Teknologi dan Informasia ASIA (JITIKA)*, 47-55.
4. Gorunescu, F. (2011). *Data Mining : Concepts, Models and Techniques*. Berlin: Springer.
5. Hand, D. J., & Yu, K. (2001). Idiot's Bayes: Not So Stupid after All? *Revue Internationale de Statistique, Vol. 69, No. 3*, 385-398.
6. Karim, M., & Rahman, R. (2013). Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. *Journal of Software Engineering and Applications*, 196-206
7. Mackay, D. J. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.
8. Murty, M. N., & Devi, V. S. (2011). *Pattern Recognition: An Algorithmic Approach*. London: Springer.
9. Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (1998). Learning to Classify Text from Labeled and Unlabeled Documents. *American Association for Artificial Intelligence*.
10. Norvig, P., & Russel, S. J. (2003). *Artificial Intelligence: A Modern Approach*. Prentice Hall.