Analisis Penerapan Algoritma *Naive Bayes* dalam Pengklasifikasian Konten Berita Bahasa Indonesia

Vipy Wahyu Perdana[1], Heru Agus Santoso[2]
Teknik Informatika-S1, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro
Jl. Nakula I No. 5-11, Semarang, 50131, (024) 3517261
email : vipywahyu@gmail.com, herezadi@gmail.com

Abstrak

Dalam tahapan proses pemuatan artikel berita, artikel yang akan diterbitkan harus diserahkan terlebih dahulu oleh editor untuk diedit dan kemudian dikategorikan sesuai isi beritanya. Untuk memudahkan kinerja editor tersebut, kita dapat memanfaatkan teknologi untuk membantu kinerja editor, yaitu dengan mengkategorikan artikel beriita secara otomatis dengan menggunakan komputer. Namun untuk melakukan hal tersebut, sistem komputer tidak bisa mengenali isi dari dokumen berita tersebut secara langsung. Diperlukan langkah khusus dalam perancangan sistem ini, agar sistem dapat mengenali keterkaitan kata – kata yang terkandung di dalam dokumen berita tersebut sehingga sistem dapat mengkategorikan berita tersebut secara tepat. Text mining merupakan salah satu metode dalam mencari informasi yang terdapat pada teks yang terdapat pada suatu dokumen. Penelitian ini menggunakan Tf-Idf dan Naive Bayes dalam mengklasifikasikan berita sebaynyak 900 dokumen yang terbagi dalam 6 kategori yaitu ekonomi & bisnis, olahraga, otomotif, politik, sosial budaya, dan teknologi. Hasil penelitian ini adalah Tf-Idf dan Naive Bayes mampu melakukan proses klasifikasi berita dengan nilai akurasi sebesar 89.22%.

Kata kunci : Klasifikasi,Kategori Berita, Text Mining, Tf-Idf, Naive Bayes

Abstract

In this stage of the process of loading of the news article, the article will be published must be submitted in advance by the editor to edit and then categorized according to its news content. To facilitate the performance of these editors, we can take advantage of technology to help the performance of editors, that is by categorizing the article beriita automatically by using a computer. But to do so, the computer system can not recognize the contents of the document the news directly. Special measures are needed in the design of this system, so that the system can recognize the inter-connection between the word – the word is contained in document the news so that the system can categorize the news accurately. Text mining is one of the methods in the search for information contained in the text which is contained in a document. TF-IDF is a step to give weight to a Word document's relationship with the concept of combining two calculations i.e. frequencies of words and the inverse of the frequency of documents containing those words. Naive Bayes algorithm is a classification algorithm, one of which is a simple probabilistic-based prediction techniques based on the application of Bayes's theorem assuming a strong ketidaktergantungan. This research uses the 900 news are divided into 6 categories, namely economic, business, sports, & automotive, socio-cultural, political, and technology. The results of this research are the Tf-Idf and Naive Bayes classification process was able to do the news with accuracy values amounted to 89.22%.

Keywords : Classification, News Category, Text Mining, Tf-Idf, Naive Bayes

1. PENDAHULUAN

1.1 Latar Belakang

Berita telah menjadi kebutuhan tersendiri bagi masyarakat. Dan dengan adanya portal berita, jumlah berita yang dihasilkan juga semakin besar. Dalam upaya untuk memudahkan kinerja editor dalam memilah dan mengkategorikan berita, saat ini kita dapat memanfaatkan teknologi untuk membantu kinerja editor tersebut, salah satunya dengan mengkategorian artikel berita secara otomatis. Namun untuk melakukan hal tersebut, sistem komputer tidak bisa melakukannya secara langsung[1]. Untuk melakukan hal tersebut, sistem harus mengenali isi dari dokumen berita tersebut. Selanjutnya mengenali hubungan antar kata dalam kalimat dan hubungan kalimat yang satu dengan yang lain serta paragraf yang satu dengan yang lainnya [2]. Oleh karena itu, diperlukan metode khusus dalam perancangan sistem ini, agar sistem dapat mengenali keterkaitan kata – kata yang terkandung di dalam dokumen berita tersebut sehingga sistem dapat mengkategorikan berita tersebut secara tepat. *Text mining* merupakan salah satu metode dalam mencari informasi yang terdapat pada teks yang terdapat pada suatu dokumen. Text mining dapat memberikan solusi atas masalah dalam memproses, mengorganisasi, menganalisa teks yang tidak terstruktur dalam jumlah besar [2]. *Term Frequency – Inversed Document Frequen*cy (TF-IDF) merupakan suatu cara untuk memberikan bobot hubungan suatu kata terhadap dokumen. Metode ini menggabungkan dua konsep untuk perhitungan bobot yaitu, frekuensi kemunculan sebuah kata didalam sebuah dokumen tertentu dan *inverse* (kebalikan) dari frekuensi dokumen yang mengandung kata tersebut [6]. *Naïve Bayes* merupakan salah satu algoritma klasifikasi yang merupakan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes dengan asumsi ketidaktergantungan yang kuat [7]. Algoritma *Naive Bayes* akan mencari nilai probabilitas kata – kata yang ada pada setiap kategori berita dan juga probabilitas dari masing – masing kategori.

1.2 Rumusan Masalah

Bagaimana langkah – langkah yang tepat dalam menerapkan *tf-idf* dan *Naive Bayes* ke dalam sistem pengklasifikasian berita ini. Sehingga sistem dapat mengklasifikasikan dokumen berita ke dalam kategori yang tepat.

1.3 Tujuan Penelitian

untuk menganalisis pengklasifikasian kategori berita ini dengan menerapkan *Term Frequency-Inverse Document Frequency* (*Tf-Idf*) sebagai pembobotan kata dan menggunakan *Naive Bayes* sebagai metode pengklasifikasiannya.

2. METODE PENELITIAN

2.1 Tinjauan Studi

Peneliti	Judul	Hasil
Amir Hamzah	Klasifikasi Teks Dengan Naive Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis	Naive Bayes dapat mengklasifikasikan dokumen berita dan akademik dengan hasil akurasi maksimal mencapai 91% pada dokumen berita, sedangkan pada dokumen akademik sebesar 82%.
Herman dkk	Klasifikasi Dokumen Naskah Dinas Menggunakan Algoritma Term Frequency – Inversed Document Frequency dan Vector Space Model	perbedaan jumlah data training mempengaruhi akurasi klasifikasi dokumen. Hasil pengujian akurasi terhadap 50 dokumen uji yang sudah dilatih sebelumnya adalah seluruh klasifikasi 100% benar. Sedangkan untuk hasil dari 50 dokumen uji yang belum pernah dilatih sebelumnya, mendapatkan akurasi berkisar antara 70 – 80 %

2.2 *Text Mining*

Text mining adalah bidang interdisipliner yang mengacu pada pencarian informasi, pertambangan data, pembelajaran mesin, statistik, dan komputasi linguistik. Langkah – langkah yang dilakukan pada proses *text mining* antara lain *text preprocessing*, yang berguna untuk mempersiapkan teks menjadi data yang akan diolah lebih lanjut. Tindakan yang dilakukan pada tahap ini adalah *to lower case*, yaitu mengubah semua karakter huruf menjadi huruf kecil, dan *tokenizing* yaitu proses penguraian deskripsi yang semula berupa kalimat – kalimat menjadi kumpulan kata dan menghilangkan spasi, karakter angka, dan delimiter - delimiter seperti tanda titik(.), koma(,) yang ada pada kata tersebut. Selanjutnya adalah f*eature selection*. Pada tahap ini tindakan yang dilakukan adalah menghilangkan *stopword* (*stopword removal*), yaitu kosakata yang bukan merupakan ciri (kata unik) dari suatu dokumen. Misalnya “di”, “oleh”, “pada”, “sebuah”, “karena”, dan lain sebagainya[3][8].

2.3 *Term Frequency – Inverse Document Frequency*(*Tf-Idf*)

Tf-Idf adalah suatu cara dalam memberikan pembobotan kata yang didasarkan pada nilai statistik yang menunjukkan kemunculan suatu kata di dalam dokumen. Pembobotan dilakukan dengan menghitung frekuensi kemunculan *term* dalam dokumen. Frekuensi kemunculan

kata (*term frequency*) merupakan petunjuk sejauh mana kata (*term*) tersebut mewakili isi dokumen. Semakin besar kemunculan suatu *term* dalam dokumen akan memberikan nilai kesesuian yang semakin besar.. Rumus perhitungan *tf-idf* adalah:

$$\text{TF-IDF}(w,d)=\text{TF}(w,d) * \left(\log\left(\frac{N}{DF(w)}\right) \right)$$

Keterangan:

TF-IDF(w,d): bobot suatu kata dalam keseluruhan dokumen
w: suatu kata (word)
d: suatu dokumen (document)
TF(w,d): frekuensi kemunculan sebuah kata w dalam dokumen d
IDF(w) inverse DF dari kata w
N: jumlah keseluruhan dokumen
DF(w): jumlah dokumen yang mengandung kata w

2.4 *Naive Bayes*

Naive Bayes merupakan salah satu algoritma yang digunakan dalam pengklasifikasian dokumen. *Naive Bayes* menerapkan teknik prediksi berbasis probabilistik sederhana yang berdasar pada penerapan teorema Bayes dengan asumsi ketidaktergantungan yang kuat. Ada beberapa tahap dalam perhitungan menggunakan *Naive* Bayes. Yang pertama dengan mencari nilai rata – rata (μ) dan standar deviasi (σ) pada data numerik, dengan persamaan:

$$\mu = \frac{\sum_{i=1}^{n} Xi}{n}$$

Keterangan:
X_i = data ke-i
n = jumlah data yang ada

Sedangakan untuk mencari standar deviasi, digunakan persamaan berikut:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(Xi - \mu)2}{n-1}$$

Keterangan :
X_i = data ke-i
μ = nilai rata – rata pada suatu data
n = jumlah data yang ada

Kemudian tahap selanjutnya adalah dengan menghitung probabilitas kepadatan dari data uji dengan rumus:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Keterangan:
x = Nilai dari data yang diujikan
σ = Standar deviasi pada data uji
μ = Rata-rata pada data uji

Selanjutnya cari nilai probabilitas dari atribut nominalnya.

$$\text{P (nominal)} = \frac{\sum_{i=1}^{n} X}{n}$$

Keterangan :
X = Nilai dari data yang diujikan
n = jumlah data yang ada

Dan selanjutnya cari nilai probabilitas dari masing – masing kategori, yaitu:

*P(kategori)=P(nominal)*P(kepadatan*

2.5 Sistem Temu Kembali Informasi

Sistem temu kembali informasi adalah pengorganisasian dan penemuan informasi dari sejumlah besar dokumen berbasis teks [3]. Sistem temu kembali informasi digunakan untuk menemukan kembali informasi yang relevan terhadap kebutuhan pengguna dari suatu kumpulan informasi secara otomatis. Tujuan dari sistem temu kembali informasi yang ideal adalah :

1. Menemukan seluruh dokumen yang relevan terhadap suatu query.
2. Hanya menemukan dokumen relevan saja, artinya tidak terdapat dokumen yang tidak relevan pada dokumen hasil pencarian.

Dua keadaan tersebut digunakan untuk menghitung performansi sistem temu kembali, yaitu *recall* dan *precision*. *Recall* dinyatakan sebagai bagian dari dokumen relevan dalam dokumen yang ditemukan, sedangkan *precision* dinyatakan sebagai bagian dokumen relevan. Persamaannya adalah sebagai berikut:

$$recall = \frac{jumlah\ dokumen\ relevan\ yang\ ditemukan}{jumlah\ semua\ dokumen\ yang\ relevan}$$

$$precision = \frac{jumlah\ dokumen\ relevan\ yang\ ditemukan}{jumlah\ semua\ dokumen\ yang\ ditemukan}$$

Keduanya menggambarkan performansi dari sistem temu kembali informasi dengan melakukan perhitungan terhadap jumlah dokumen relevan hasil pencarian. Pengukuran *recall* dan *precision* ini merupakan perhitungan yang dilakukan terhadap kumpulan dokumen hasil pencarian (*set based measure*) secara keseluruhan. Pengukuran dengan menggunakan *set based measure* ini tidak dapat menggambarkan performansi sistem temu kembali informasi mengenai urutan dari dokumen-dokumen relevan [5].

3. HASIL PEMMODELAN

Hasil yang didapat dari penelitian ini adalah *Tf-idf* dan *Naive Bayes* mampu melakukan proses klasifikasi 6 kategori berita ini dengan baik. Pada kategori ekonomi & bisnis, dokumen yang masuk dalam kategori yang tepat berjumlah 128 dokumen. Pada kategori olahraga, dokumen yang masuk dalam kategori yang tepat berjumlah 141 dokumen. Pada kategori otomotif, dokumen yang masuk dalam kategori yang tepat berjumlah 141 dokumen. Pada kategori politik, dokumen yang masuk dalam kategori yang tepat berjumlah 130 dokumen. Pada kategori sosial budaya, dokumen yang masuk dalam kategori yang tepat berjumlah 124 dokumen. Dan pada kategori teknologi, dokumen yang masuk dalam kategori yang tepat berjumlah 139 dokumen. Rata – rata nilai *recall* yang dihasilkan sebesar 89.23%, dan rata – rata nilai *precision* sebesar 89.44%. Serta nilai akurasi dari klasifikasi ini mencapai 89.22%.

4. KESIMPULAN DAN SARAN

4.1 Kesimpulan

Dari penelitian klasifikasi kategori berita ini dapat disimpulkan bahwa penggunaan *Tf-Idf* dan *Naive Bayes* dapat digunakan untuk proses klasifikasi berita dengan nilai akurasi sebesar 89.22%.

4.2 Saran

1. Menambahkan proses *stemming* dalam proses pengolahan kata. Sehingga kata – kata yang masih mempunyai imbuhan dapat diubah menjadi kata dasar, dan dapat disesuaikan dengan daftar kata pada *stopword list*. Sehingga dapat digunakan sebagai kata kunci pada suatu dokumen.
2. Menambah jumlah data sebanyak mungkin, sehingga hasil klasifikasi dapat ditingkatkan lagi.

DAFTAR PUSTAKA

[1] Agustoni and Fitri M Sari, "Pemilihan Artikel Berita dengan Text Mining," Oktober 2011.

[2] Wiwin Sulistyo, "Klasifikasi Dokumen Berbahasa Inggris Berdasarkan Weighted-Term," JURNAL DUNIA TEKNOLOGI INFORMASI Vol. 1, No. 1, (2012) 14-19, 2008.

[3] Amalia Indranandita, Budi Susanto, and Antonius Rachmat C, "Sistem Klasifikasi dan Pencarian Jurnal dengan Menggunakan Metode Naive Bayes dan Vector Space Model," JURNAL INFORMATIKA, VOLUME 4 NOMOR 2, November 2OO8.

[4] Prajna W Basnur and Dana I Sensue, "Pengklasifikasian Otomatis Berbasis Ontologi untuk Artikel Berita Berbahasa Indonesia," Makara, Teknologi, Vol. 14, No. 2, April 2010: 29-35, 2010.

[5] Rila Mandala and Hendra Setiawan, "Peningkatan Performansi Sistem Temu-

Kembali dengan Perluasan Query Secara Otomatis," Bandung : Departemen Teknik Informatika Institut Teknologi Bandung, 2002.

[6] Rolly Intan and Andrew Defeng, "Hard: Subject-Based Engine menggunakan TF-IDF dan Jaccard's Coefficient ," JURNAL TEKNIK INDUSTRI VOL. 8, NO. 1, JUNI 2006: 61-72, 2006.

[7] Prasetyo Eko, Data Mining : Konsep dan Aplikasi menggunakan Matlab. Jogjakarta: Penerbit Andi, 2012.

[8] Bambang Kurniawan, Syahril Effendi, and Opim Salim Sitompul, "Klasifikasi Konten Berita Dengan Metode Text Mining," JURNAL DUNIA TEKNOLOGI INFORMASI Vol. 1, No. 1, (2012) 14-19, 2012.

[9] Amir Hamzah, "Klasifikasi Teks dengan Naive Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis," Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) Periode III, November 2012.

[10] Herman, Andani Achmad, and Amil Ahmad Ilham, "Klasifikasi Dokumen Naskah Dinas Menggunakan Algoritma Term Frequency-Inversed Document Frequency dan Vector Space Model," Balai Besar Pengkajian dan Pengembangan Komunikasi dan Informatika Makassar,Kementerian Komunikasi dan Informatika, 2012.

[11] Danang T Massandy and Masayu L Khodra, "Klasifikasi Kategori Berita dengan Metode Pembelajaran Semi Supervised," 2012.

[12] Oka Karmayasa and Ida B Mahendra, "Implementasi Vector Space Model dan Beberapa Notasi Metode Term Frequenct Inverse Document Frequency (TF-IDF) pada Sistem Temu Kembali Informasi," 2010.

[13] Agus Z Arifin and Ari N Setiono, "Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia," 2002.

[14] Lanang Febria Galing Gumilang, "PROSES KERJA REPORTER BERITA DETIKHOT SUBKANAL MUSIC DI DETIK.COM," 2009.

[15] Ni Wayan Sumartini Saraswati, "Text Mining dengan Metode Naive Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis," Universitas Udayana : Indonesia, 2011.

[16] Ian H.Witten, "Text mining," University of Waikato, 2003.