Implementasi Naïve bayes Clasifier dalam Klasifikasi Jenis Berita

Dessy Santi[1], Jumadil Nangi [2] Natalis Ransi [3]
[1)], Fakultas Teknik, Jurusan Teknologi Informasi, Universitas Tadulako, Palu
e-mail: dessy.santi81@gmail.com
[2,3)]Jurusan Teknik Informatika, Fakultas Teknik, Universitas Halu Oleo, Kendari
e-mail: *[2]jumadilnangi87@gmail.com, [3] natalis.ransi@gmail.com

Abstract

Sometimes the classification of news categories is still an obstacle. Classification can be wrong because it is still subjective. As a result, the selected category does not match the uploaded news description. Based on these problems, the authors feel the need to make Classification of News Types with the Naïve Bayes Classifier Algorithm. The importance of this system is to be able to classify news and help news seekers to get the news they want.

Based on the test results, the Naïve Bayes Classifier algorithm has a good performance for the classification of news types. This is evidenced in testing using news data taken from www.kompasiana.com, then news is classified into four categories namely politics, economics, sports, and entertainment. The classification results using 16 test news obtained an accuracy of 87.5%.

Kata kunci— Portal Berita, *Text Mining*, *Naïve Bayes Classifier*.

I. PENDAHULUAN

Berita merupakan informasi baru atau informasi mengenai sesuatu yang sedang terjadi, disajikan lewat bentuk cetak, siaran, *internet*, atau dari mulut ke mulut kepada orang ketiga atau orang banyak. Berdasarkan kamus besar bahasa Indonesia, yang dimaksud dengan berita adalah cerita atau keterangan yang terdiri dari suatu kejadian atau peristiwa yang baru. Berita juga bisa disebut juga dengan cerita atau keterangan mengenai kejadian atau peristiwa yang hangat.

Pada era perkembangan teknologi saat ini, berita dapat dilihat menggunakan internet seperti www.kompasiana.com yang merupakan salah satu website berita yang sering dikunjungi. Banyak informasi yang dapat kita terima dalam *website* tersebut. Terkadang pengklasifikasian kategori berita masih menjadi kendala. Pengklasifikasian bisa saja salah karena masih subjektif.

Untuk mempermudah dalam pengklasifikasian kategori berita, diperlukan sebuah sistem dengan menggunakan metode text mining sebagai salah satu alternatif untuk menyelesaikannya. Text mining adalah salah satu teknik yang dapat digunakan untuk melakukan klasifikasi, dimana text mining merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar. Selain klasifikasi, text mining juga digunakan untuk menangani masalah clustering, information extraction, dan information retrival.

Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifier, seperti Decision Tree Classifie dan Neural Network Classifier. Naïve Bayes Classifier memiliki tingkat akurasi 95.20% lebih baik dibanding model classifier Decision Tree Classifier dan Neural Network Classifier. Keuntungan penggunaan Naïve Bayes Classifier adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian [1].

Melihat permasalahan di atas, maka penulis mengembangkan sebuah sistem yang mampu mengklasifikasikan berita

secara otomatis. Atas dasar inilah penulis mengangkat judul "Implementasi Text Mining untuk Klasifikasi Jenis Berita menggunakan Algoritma Naïve Bayes Classifier."

II. METODE PENELITIAN

2.1. *Portal Berita*

portal berita adalah situs yang menampilkan informasi mengenai informasi yang terjadi ke masyarakat. Keberadaan portal berita tidak terlepas dari segala hal yang berhubungan dengan berita, seperti jenis berita bagian berita, dan unsur berita.

2.2. *Kompasiana.com*

Kompasiana adalah sebuah situs website yang kontennya berasal dan dibuat oleh warga. Setiap orang yang berkontribusi di Kompasiana harus terdaftar sebagai anggota atau Kompasianer sebutan akrab anggota Kompasiana. Semua konten yang tayang di Kompasiana dikelola oleh Kompasianer (atau lazim disebut *User Generated-Content*).

Pada tahun 2008, Kompasiana merupakan blog jurnalis dan karyawan di lingkungan Kompas Gramedia. Nama Kompasiana diusulkan oleh Budiarto Shambazy, wartawan senior harian KOMPAS. Nama ini pernah digunakan untuk kolom khusus yang dibuat pendiri harian KOMPAS, PK Ojong, berisi tulisan mengenai situasi mutahir pada masanya. Kumpulan rubrik Kompasiana yang ditulis PK Ojong kini sudah dibukukan. Pada September 2008, Kompasiana bertransformasi sebagai blog sosial atau media warga dengan kontribusi sekitar 300.000 anggota yang tersebar di berbagai penjuru dunia. Anggota Kompasiana berasal dari berbagai macam latar belakang profesi, usia dan pendidikan.

Perharinya, Kompasiana menayangkan sekitar 400-700 artikel berupa laporan, opini dan karya fiksi. Tampilan *website* Kompasiana sudah beberapa kali mengalami perubahan antarmuka. Yang paling terbaru, Kompasiana memiliki tampilan yang bersih, ringan diakses dan lebih interaktif dengan hadirnya fitur "obrolan". Saat ini Kompasiana telah menayangkan lebih dari 2 juta artikel dengan kunjungan pembaca di tiap bulannya sebesar 18 juta pengunjung. Kompasiana juga masuk ke dalam 10 besar *website* berita di Indonesia dan 5 besar *website* buatan anak bangsa.

Semua konten yang tayang di Kompasiana dikelola oleh Kompasianer atau *User*. *User* dapat mengunggah berita dan memilih kategori yang diinginkan. Kompasiana belum memiliki fitur klasifikasi kategori berita otomatis dalam pengunggahan berita. Oleh karena itu, penulis memilih Kompasiana sebagai *website* acuan dalam penelitian ini [2].

2.3. *Data Mining*

Data mining adalah serangkaian proses untuk menggali nilai tambah berupa informasi yang selama ini tidak diketahui secara manual dari suatu basis data dengan melakukan penggalian pola-pola dari data dengan tujuan untuk memanipulasi data menjadi informasi yang lebih berharga yang diperoleh dengan cara mengekstraksi dan mengenali pola yang penting atau menarik dari data yang terdapat dalam basis data.

Knowledge Discovery in Database (KDD) merupakan proses pencarian pengetahuan yang bermanfaat dari kumpulan data. Proses KDD bersifat interaktif dan iteratif, meliputi sejumlah langkah dengan melibatkan pengguna dalam membuat keputusan dan dapat dilakukan pengulangan di antara dua buah langkah. Data mining merupakan salah satu proses inti yang terdapat dalam *Knowledge Data Discovery* (KDD). Banyak orang memperlakukan data mining sebagai sinonim dari KDD, karena sebagian besar pekerjaan dalam KDD difokuskan pada data mining.Namun, langkah-langkah ini merupakan proses yang penting yang menjamin kesuksesan dari aplikasi KDD [3].

2.4. *Text Mining*

Text mining merupakan variasi dari *Data mining* yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yang berjumlah besar. *Text mining* didefinisikan sebagai data yang berupa teks yang biasanya sumber data didapatkan dari dokumen, dengan tujuan adalah mencari kata-kata yang dapat mewakili isi dari dokumen tersebut yang nantinya dapat dilakukan analisa hubungan antar dokumen [4].

Tahapan *text mining* secara umum dibagi menjadi beberapa tahapan umum [5].

2.5. *Text Preprocessing*

Text Preprocessing merupakan tahapan awal dari *text mining* yang bertujuan mempersiapkan teks menjadi data yang akan mengalami pengolahan pada tahap selanjutnya. Pada *text mining*, data mentah yang berisi informasi memiliki struktur yang sembarang, sehingga diperlukan proses pengubahan bentuk menjadi data yang terstruktur sesuai kebutuhan, yaitu biasanya akan mejadi nilai-nilai numerik. Proses ini disebut *Text Preprocessing*.

Pada tahap ini, tindakan yang dilakukan adalah *toLowerCase*, dengan mengubah semua karakter huruf menjadi huruf kecil, dan *tokenizing* yaitu proses penguraian deskripsi yang semula berupa kalimat mejadi kata-kata kemudian menghilangkan delimiter-delimiter seperti tanda koma (,), tanda titik (.), spasi, dan karakter angka yang terdapat pada kata tersebut.

2.6. *Seleksi Fitur (Feature Selection)*

Pada tahap ini akan dilakukan seleksi dengan mengurangi jumlah kata-kata yang dianggap tidak penting dalam dokumen tersebut untuk menghasilkan proses pengklasifikasian yang lebih efektif dan akurat. Tahapan ini adalah dengan melakukan penghilangan *stopword* dan juga mengubah kata-kata kedalam bentuk dasar terhadap kata yang berimbuhan.

Stopword merupakan kosakata yang bukan merupakan ciri atau kata unik dari suatu dokumen seperti kata sambung. Yang termasuk *stopword* yaitu " di", "pada", "sebuah", "karena", "oleh" dan sebagainya. Sebelum memasuki tahapan penghilang *stopword*, daftar *stopword* harus dibuat terlebih dahulu. Jika kata-kata yang termasuk *stopword* masuk dalam *stoplist*, maka kata tersebut akan dihapus dari deskripsi sehingga sisanya dianggap sebagai kata-kata yang mencirikan isi dokumen atau *keywords.*

2.7. *Stemming*

Stemming adalah proses pemetaan dan penguraian berbagai bentuk dari suatu kata menjadi kata dasarnya. Tujuan dilakukannya proses *stemming* adalah menghilangkan imbuhan-imbuhan berupa *prefix*, *suffix*, maupun konfiks yang terdapat pada setiap kata. Apabila imbuhan tadi tidak dihilangkan maka setiap kata akan disimpan di dalam *database*, sehingga nantinya akan menjadi beban di dalam *database*. Bahasa Indonesia memiliki aturan morfologi maka proses *stemming* harus berdasarkan aturan morfologi bahasa Indonesia.

2.8. *Naïve Bayes Classifier*

Algoritma *Naïve Bayes Classifier* merupakan algoritma yang digunakan untuk mencari nilai probabilitas tertinggi untuk mengklasifikasi data uji pada kategori yang paling tepat. Dalam penelitian ini yang menjadi data uji adalah dokumen berita. Ada dua tahap pada klasifikasi dokumen. Tahap pertama adalah pelatihan terhadap dokumen yang sudah diketahui kategorinya. Sedangkan tahap kedua adalah proses klasifikasi dokumen yang belum diketahui kategorinya [6].

Dalam penelitian ini, penulis melakukan pengumpulan data dengan cara yaitu :

2.9. *Studi Kepustakaan*

Metode pengumpulan data yang digunakan dalam penelitian ini adalah dengan

melakukan studi kepustakaan. Melalui studi pustaka penulis menghimpun data dari jurnal dan *www.kompasiana.com* yang berkaitan dengan pembangunan sistem implementasi text mining untuk klasifikasi jenis berita menggunakan *Naïve Bayes Classifier*. Baik itu untuk perancangan basis data, antarmuka dan operasi standar.
Dalam metode ini terdapat beberapa tahapan, yaitu:

2.10. *Inception*

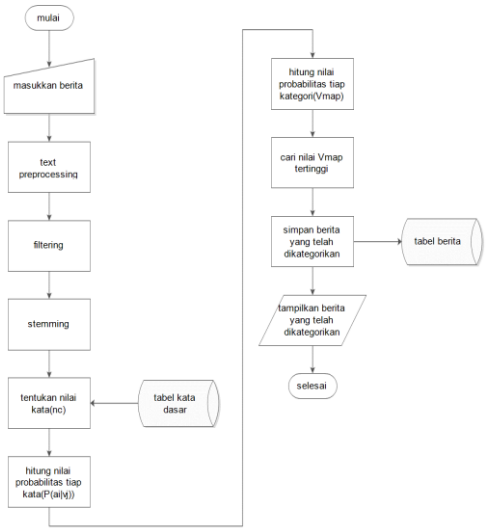
Pada tahap ini dilakukan penentuan ruang lingkup dan kebutuhan sistem. Ruang lingkup dan kebutuhan disesuaikan dengan hasil yang diperoleh saat pengumpulan data. Adapun hasil yang diperoleh berkaitan dengan sistem ini adalah sistem mengolah data yang berkaitan dengan sistem Implementasi *Text Mining* untuk Klasifikasi Jenis Berita menggunakan Algoritma *Naïve Bayes Classifier*.

2.11. *Elaboration*

Perencanaan dari Implementasi *Text Mining* untuk Klasifikasi Jenis Berita menggunakan Algoritma *Naïve Bayes Classifier* dilakukan di tahap ini. Meliputi rancangan antarmuka, rancangan basis data dan seleksi operasi apa saja yang dapat diimplementasikan sesuai dengan ruang lingkup dan kebutuhan yang telah didefinisikan pada tahap permulaan (*inception*). Adapun hasil yang dapat diperoleh pada tahap perluasan adalah sebagai berikut:

a. *Website* yang akan dibangun terbagi menjadi dua *level* pengguna yaitu, *user* dan *admin*.
b. *User* hanya dapat melihat berita terbaru pada halaman beranda sistem untuk *user*.
c. *Website* yang dibangun akan secara otomatis mengklasifikasikan berita ke dalam kategori-kategori yang sudah ditentukan.

Adapun hasil rancangan flowchart diagram Sistem Implementasi Text Mining untuk Klasifikasi Jenis Berita menggunakan Algoritma Naïve Bayes Classifier ditunjukkan oleh Gambar 1.



Gambar 1. *Flowchart Diagram Sistem*

2.12. *Construction*

Pada tahap ini dilakukan pembangunan Sistem Implementasi *Text Mining* untuk Klasifikasi Jenis Berita menggunakan Algoritma *Naïve Bayes Classifier* melalui proses penulisan kode program (*coding*), pembuatan basis data, dan penginputan data. Pembangunan sistem dilakukan berdasarkan rancangan yang telah direncanakan pada tahap perluasan/ perencanaan (*elaboration*).

2.13. *Transition*

Pada tahap ini dilakukan proses pengujian terhadap Implementasi *Text Mining* untuk Klasifikasi Jenis Berita menggunakan Algoritma *Naïve Bayes Classifier*.

III. HASIL DAN PEMBAHASAN

3.1. *Pengujian Hasil Klasifikasi*

Adapun pengujian sistem, dilakukan dengan Pengujian hasil klasifikasi dilakukan untuk mengetahui tingkat keakurasian sistem implementasi *text mining* untuk klasifikasi jenis berita menggunakan Algoritma *Naïve Bayes Classifier*. Pengujian dilakukan pada hasil kelas untuk data uji. Kategori berdasarkan

portal berita online yang sudah ada (*website www.kompasiana.com*) disimbolkan dengan KPB, dan kategori berdasarkan sistem yang telah dibuat disimbolkan KSU.

Hasil model pengujian untuk mengetahui hasil klasifikasi serta mencocokkan kategori berita uji yang ada *www.kompasiana.com* dengan sistem yang telah dibuat ditunjukkan oleh Tabel 1.3

Tabel 1 Hasil Klasifikasi Sistem Implementasi *Text Mining* untuk Klasifikasi Jenis Berita Menggunakan Algoritma *Naïve Bayes Classifier*

Subset	No	Judul Berita	KPB	KSU	Pengujian
1	1	Risma Bersedia ke DKI I	Politik	Politik	Berhasil
	2	RUU Dwi Kewarganegaraan	Politik	Politik	Berhasil
	3	Demokrasi Sekarang Menurut Rakyat Indonesia	Politik	Politik	Berhasil
	4	Ahok, Oh Ahok…	Politik	Politik	Berhasil
2	5	Gubernur Bersih Versi BPK Menjadi Tersangka KPK	Ekonomi	Politik	Gagal
	6	Sesat Pikir Wapres "Bunga Deposito Harus Turun"	Ekonomi	Ekonomi	Berhasil
	7	Wacana Kenaikan Cukai yang Berimbas Terhadap Naiknya Harga Rokok	Ekonomi	Ekonomi	Berhasil
	8	Harga Rokok Naik 50rb /Bungkus, ini yang akan Terjadi pada Indonesia	Ekonomi	Ekonomi	Berhasil
3	9	Hasil Undian Liga Champion 2016/17	Olahraga	Olahraga	Berhasil
	10	Indonesia Apresiasi Terhadap ATLET Peraih Olimpiade Rio de Janerio	Olahraga	Olahraga	Berhasil
	11	Tersisa Sekeping Sejarah di Rio 2016	Olahraga	Olahraga	Berhasil
	12	2 Karateka Puteri Kodam Jaya Sumbang 2 Medali Perunggu	Olahraga	Olahraga	Berhasil
4	13	Jangan Tenggelamkan Lagu Anak Indonesia	*Entertainment*	*Entertainment*	Berhasil
	14	Retta Sitorus Semakin Populer di Blantika Musik Batak	*Entertainment*	*Entertainment*	Berhasil
	15	Erick Sihotang Artis Batak Top 2014-2016	*Entertainment*	*Entertainment*	Berhasil
	16	Demam Mukidi	*Entertainment*	Politik	Gagal

IV. KESIMPULAN

Berdasarkan pembahasan dan evaluasi dari bab sebelumnya, maka kesimpulan yang dapat diambil adalah sebagai berikut :

1. Kategori berita yang digunakan adalah empat kategori, yaitu berita politik, ekonomi, olahraga, dan *entertainment* dimana data berita tersebut diambil dari media berita *online*.
2. Berita yang digunakan dalam penelitian ini hanya berita berbahasa Indonesia.
3. Berita yang telah melalui proses pengklasifikasian pada sistem, bisa memiliki lebih dari satu kategori.
4. Lama waktu proses pengklasifikasian kategori berita tergantung dari banyaknya kata yang terkandung dalam isi berita.
5. Algoritma *Naïve Bayes Classifier* memiliki kinerja yang baik untuk klasifikasi jenis berita. Hal ini dibuktikan pada pengujian menggunakan data berita yang diambil dari *www.kompasiana.com*, kemudian berita diklasifikasikan pada empat kategori yaitu politik, ekonomi, olahraga, dan *entertainment*. Hasil klasifikasi menggunakan 16 berita uji didapatkan akurasi 87.5%.

DAFTAR PUSTAKA

[1] Xhemali, D., Hinde, C.J., Stone, R.G. 2009. *Naïve Bayes vs. Decision Trees vs. Neural Networks in the classification of training web page*s. Loughborough University, Loughborough.

[2] Anonim. 2013., Profile, http://www.kompasiana.com/tentang-kompasiana. Diakses tanggal 15 September 2016.

[3] Tan, Pang-Ning. 2006. *Introduction to Data Mining.* USA : Pearson Addison Wesley.

[4] Milka, H. 2006. *Machine Learning Text Kategorization*. Austin : University of Tex.

[5] Triawati, C. 2009. *Metode Pembobotan Statistical Concept Based untuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia*. Institut Teknologi Telkom, Bandung.

[6] Kurniawan, B., Effendi, S., Sitompu, O.S. 2012. Klasifikasi Konten Berita Dengan Metode Text Mining. Universitas Sumatera Utara, Medan.