KLASIFIKASI ARTIKEL BERITA MENGGUNAKAN METODE *TEXT MINING* DAN *NAIVE BAYES CLASSIFIER*

Ira Anggraeni Setiawan [1)], Tacbir Hendro P [2)], Dian Nursantika [3)]

[1),2),3)]Informatika, Universitas Jenderal Achmad Yani
Jl. Terusan Jenderal Sudirman Po Box 148 Cimahi
Email : iranggraeni15@gmail.com

Abstrak. *Perkembangan teknologi informasi dan komunikasi memang cukup menakjubkan. Kecenderungan seseorang untuk mengakses informasi khususnya berita melalui dunia maya pun menjadi semakin tinggi. Informasi merupakan hal yang sangat terpenting dalam kehidupan bermasyarakat. Salah satu sumber informasi adalah melalui web portal atau website berita. Ada sekitar 300 sampai 400 artikel berita dalam satu bulan dan banyaknya kategori artikel atau kanal atau rubrik dalam sebuah web portal, membuat kinerja editor semakin banyak karena di sini seorang editor harus dapat mengedit artikel berita dari berbagai kanal dan sekaligus harus mengkategorikan artikel satu persatu secara manual ke dalam beberapa kategori yang ditentukan. Oleh karena itu dalam penelitian ini akan dirancamg sistem klasifikasi yang dapat mengelompokkan artikel berita menggunakan metode text mining dan naive bayes classifier (NBC). Klasifikasi ini ditekankan untuk data artikel berbahasa indonesia. Artikel berita ini akan dikelompokkan ke dalam 9 kanal yaitu news, finance, sport, otomotif, entertaint, healty, food, travel dan teknologi. Pada sistem klasifikasi ini menggunakan proses pembelajaran dan untuk proses keterkaitan antar data artikel diukur berdasarkan nilai probabilitas dari data dan kata yang ada.*

***Kata kunci**: Sistem Klasifikasi, Artikel Berita, Pengelompokan, Text Mining, Naive Bayes Classifier.*

1. Pendahuluan

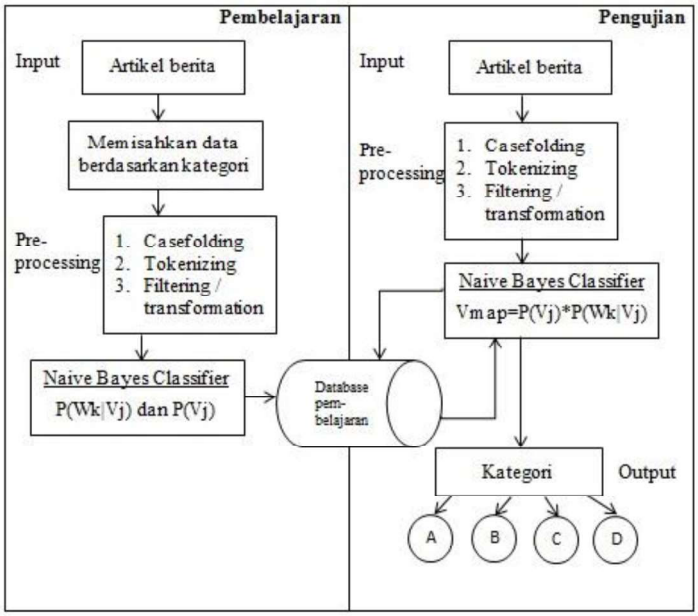
Informasi merupakan hal yang sangat terpenting dalam kehidupan bermasyarakat. Salah satu sumber informasi adalah melalui web portal atau website berita. Web portal atau website berita adalah situs yang mengumpulkan dan menyediakan aneka informasi dari berbagai sumber untuk ditampilkan kepada *user*, jika *user* tertarik untuk mengetahui informasi yang ada dengan lebih lengkap, *user* akan diarahkan ke sumber yang aslinya. Namun pada umumnya web portal tidak hanya menampilkan informasi dari sumber luar, kadang mereka juga menampilkan informasi - informasi dalam website mereka sendiri.

Perkembangan teknologi informasi dan komunikasi memang cukup menakjubkan. Kecenderungan seseorang untuk mengakses informasi khususnya berita melalui dunia maya pun menjadi semakin tinggi. Perkembangan internet sudah merambah ke berbagai lapisan masyarakat dan berbagai kalangan, mulai dari anak kecil hingga orang dewasa. Ada sekitar 300 sampai 400 artikel berita dalam satu bulan dan banyaknya kategori artikel atau kanal atau rubrik dalam sebuah web portal, membuat kinerja editor semakin banyak karena di sini seorang editor harus dapat mengedit artikel dari berbagai kanal dan sekaligus harus mengkategorikan artikel satu persatu secara manual ke dalam beberapa kategori yang ditentukan.

Untuk mengetahui kategori pada sebuah artikel berita secara otomatis tanpa harus dikategorikan dibaca satu persatu maka perlu dilakukan pengukuran kemiripan dokumen terkait dengan menggunakan metode *Naive Bayes Classifier* (NBC) yaitu proses pengenalan teks dan dokumen. Seperti salah satu kasus terdahulu menggunakan metode tersebut untuk penelitian klasifikasi pengaduan melalui web[1]. NBC adalah klasifikasi statistik yang bisa memprediksi probabilitas sebuah kelas, dan kelebihan dari metode ini adalah tingkat akurasi yang tinggi juga waktu komputasi yang lebih cepat [2].

2. Pembahasan

Sistem klasifikasi artikel berita merupakan suatu sistem yang mendukung pengelolaan data berbentuk teks secara otomatis dari hasil pengolahan data, informasi dan perancangan sistem. Dalam sistem klasifikasi artikel berita yang baru, akan menggunakan metode *naive bayes classifier* (NBC) dan proses yang paling penting dalam sistem ini yaitu penambangan sebuah teks pada suatu dokumen, sehingga dapat memberikan hasil yang sesuai dengan tujuan yang akan dicapai. Berikut ini gambaran umum sistem yang akan dibuat pada Gambar 1.



Gambar 1. Gambaran Umum Sistem Klasifikasi

Dapat dilihat pada gambar 1 gambaran umum sistem yang akan dibangun terdapat dua proses yang berbeda yaitu proses pembelajaran dan proses pengujian.

1. Pada proses pembelajaran, data artikel berita yang sudah dimasukkan akan dipisahkan berdasarkan kategori yang telah ditentukan. Kemudian data yang sudah dipisahkan masuk ketahap *preprocessing* (*casefolding, tokenizing dan filtering*). Setelah itu dihitung nilai probabilitas kata dan probabilitas kategori pada setiap data yang dijadikan data pembelajaran, kemudian disimpan ke dalam *database* pembelajaran yang berisi kata – kata penting pada setiap kategori.
2. Sedangkan pada proses pengujian yang menjadi data masukkan yaitu data artikel baru yang belum diketahui kategorinya. Pada tahap *preprocessing* yang dilakukan sama seperti yang dilakukan pada proses *preprocessing* di proses pembelajaran, yang membedakannya yaitu pada saat perhitungan probabilitas kata. Setelah melakukan tahap *preprocessing*, maka dokumen baru tersebut akan melalui proses persamaan kata. Kata – kata yang ada di data baru dengan kata – kata yang ada di *database* pembelajaran. Gunanya untuk menghitung probabilitas kata yang sama pada *database* pembelajaran. Sehingga menghasilkan probabilitas pada setiap kategori yang ada.

a. Proses Pembelajaran

Pada proses pembelajaran, data artikel yang sudah dimasukkan akan dipisahkan berdasarkan kategori, setelah itu masuk ke tahap *preprocessing* (*casefolding, tokenizing, filtering / transformation*), kemudian tahap *patern discovery* menghitung *frekuensi* banyaknya kemunculan kata, lalu menghitung nilai probabilitas. Sebagai contoh data artikel dengan judul untuk kategori news: “Kekerasan Terhadap Anak Makin Brutal, KPAI Minta Semua Pihak Turun Tangan”, untuk kategori finance: “Kementan Bantah Harga Kentang Jatuh Karena Serbuan Impor”, untuk kategori sport: “Demi Persiapan Asian Games, Dewan Olimpiade Asia Berkantor Di Jakarta”, untuk kategori otomotif: “Sama Seperti Mobil, Volume Pelumas Motor Pun Perlu Diperiksa”, untuk kategori entertain: “Sedang Sakit, Anak Ketiga

Anang Rayakan Ulang Tahun Bersama Anak Yatim", untuk kategori healty: "Salah Satu Cara Sederhana Atasi Stres Atur Pernafasan", untuk kategori food: "Ini Makanan, Minuman Hingga Resep Yang Paling Sering Dicari Di Google", untuk kategori travel: "Tak Hanya Pantai, Lombok Punya Hutan Lemor Asli" dan untuk kategori tekno: "Instagram Terbaru Bisa Simpan Postingan Orang Untuk Dilihat Nanti".

Dengan melalui proses *text mining* yaitu proses *preprocessing* (*casefolding, tokenizing dan filtering*), maka didapat jumlah kata sebanyak 43 dan jumlah frekuensi dari masing – masing kategori news = 6, jumlah frekuensi kategori finance = 5, jumlah frekuensi kategori sport = 7, jumlah frekuensi kategori otomotif = 5, jumlah frekuensi kategori entertain= 6, jumlah frekuensi kategori healty = 5, jumlah frekuensi kategori food = 3, jumlah frekuensi kategori travel = 4 dan jumlah frekuensi kategori teknologi = 3. Kemudian hitung nilai probabilitas dengan menggunakan rumus :

P(Wk|Vj) = (nk+1) / (Jml Frekuensi+Jml Kata).

Dimana :

$P(W_k|V_j)$: Probabilitas bobot kata sesuai kategori

n_k : Nilai kemunculan frekuensi kata

Sebagai contoh dari kategori finance didapat kata dari hasil preprocessing yaitu: "kementan harga kentang jatuh impor", kemudian dihitung nilai probabilitasnya menggunakan rumus di atas.

P(kementan| Finance) = (1+1) / (5+43) = 0.0416
P(harga| Finance) = (1+1) / (5+43) = 0.0416
P(kentang| Finance) = (1+1) / (5+43) = 0.0416
P(jatuh| Finance) = (1+1) / (5+43) = 0.0416
P(impor| Finance) = (1+1) / (5+43) = 0.0416

Lalu dimasukkan kedalam tabel probabilitas seperti pada Tabel 1.

Tabel 1. Probabilitas Setiap Kata

Kata	Probabilitas								
	News	Finance	Sport	Otomotif	Entertain	Healty	Food	Travel	Tekno
harga	0.0204	0.0416	0.002	0.0208	0.0204	0.0208	0.0217	0.0212	0.0217
kentang	0.0204	0.0416	0.002	0.0208	0.0204	0.0208	0.0217	0.0212	0.0217
jatuh	0.0204	0.0416	0.002	0.0208	0.0204	0.0208	0.0217	0.0212	0.0217
impor	0.0204	0.0416	0.002	0.0208	0.0204	0.0208	0.0217	0.0212	0.0217
persiapan	0.0204	0.0208	0.004	0.0208	0.0204	0.0208	0.0217	0.0212	0.0217

Setelah mendapatkan nilai probabilitas kata pada setiap kategori, kemudian hitung probabilitas kategori dengan menggunakan rumus :

P(Vj) = Jml Dokumen setiap Kategori / Total Dokumen

Diketahui : Jumlah Dokumen News =1
: Jumlah Dokumen Finance =1
: Jumlah Dokumen Sport =1
: Jumlah Dokumen Otomotif =1
: Jumlah Dokumen Entertain =1
: Jumlah Dokumen Healty =1
: Jumlah Dokumen Food =1
: Jumlah Dokumen Travel =1
: Jumlah Dokumen Teknologi =1

Jadi, probabilitas dari dokumen adalah :

P(News) = 1/9 = 0.11
P(Finance) = 1/9 = 0.11
P(Sport) = 1/9 = 0.11
P(Otomotif) = 1/9 = 0.11
P(Entertain) = 1/9 = 0.11

P(Healty) = 1/9 = 0.11
P(Food) = 1/9 = 0.11
P(Travel) = 1/9 = 0.11
P(Teknologi) = 1/9 = 0.11

b. Proses Pengujian

Contoh artikel berita dengan dokumen baru yang belum diketahui kategorinya. Contoh judul artikel berita : " Harga Kentang Impor Dipasaran Dijual Hanya Rp. 2.500 – Rp. 3.500/kg".
Setelah melalui proses text mining yaitu preprocessing, maka kata – kata yang didapat yaitu : "harga kentang impor dipasaran dijual rp rp kg". Seperti pada Tabel 2.

Tabel 2. Frekuensi Kemunculan Kata

Kata	Frekuensi
Harga	**1**
Kentang	**1**
Impor	**1**
Dipasaran	1
Dijual	1
Rp	2
Kg	1
Jumlah kata = 7	**Jumlah frekuensi = 8**

Pada tabel 2 didapat dari banyaknya frekuensi kemunculan kata pada data uji. Kemudian akan dihitung nilai probabilitas pada setiap kategori, nilai yang dimasukkan diambil dari tabel 1 pada tabel probabilitas pada proses pembelajaran.

1. Kategori News
   P(harga|News) = 0.0204
   P(kentang|News) = 0.0204
   P(impor|News) = 0.0204
   Jadi P(|News) = 0.0204*0.0204*0.0204
   = 0.000008489664
   Probabilitas = P(News) * P(|News)
   = 0.11 * 0.000008489664
   = 0.00000093386304 = **93.386.304 x 10^{-7}**
2. Kategori Finance
   P(harga| Finance) = 0.0416
   P(kentang| Finance) = 0.0416
   P(impor| Finance) = 0.0416
   Jadi P(| Finance) = 0.0416*0.0416*0.0416
   = 0.000079778816
   Probabilitas = P(Finance)* P(| Finance)
   = 0.11 * 0.000079778816
   = 0.00000877566976 = **877.566.976 x 10^{-6}**
3. Kategori Sport
   P(harga| Sport) = 0.002
   P(kentang| Sport) = 0.002
   P(impor| Sport) = 0.002
   Jadi P(| Finance) = 0.002 * 0.002 * 0.002 = 0.000000008
   Probabilitas = P(Sport) * P(|Sport)
   = 0.11 * 0.000000008
   = 0.00000000088 = **88 x 10^{-10}**
4. Kategori Otomotif
   P(harga| Otomotif) = 0.0208
   P(kentang| Otomotif) = 0.0208

P(impor| Otomotif) = 0.0208
Jadi P(| Otomotif) = 0.0208 * 0.0208 * 0.0208
= 0.000008998912
Probabilitas = P(Otomotif) * P(|Otomotif)
= 0.11 * 0.000008998912
= 0.00000098988032 = **98.988.032 10^{-7}**

5. Kategori Entertain
P(harga| Entertain) = 0.0204
P(kentang| Entertain) = 0.0204
P(impor| Entertain) = 0.0204
Jadi P(| Entertain) = 0.0204*0.0204*0.0204
= 0.000008489664
Probabilitas = P(Entertain) * P(| Entertain)
= 0.11 * 0.000008489664
= 0.00000093386304 = **93.386.304 x 10^{-7}**
6. Kategori Healty
P(harga| Healty) = 0.0208
P(kentang| Healty) = 0.0208
P(impor| Healty) = 0.0208
Jadi P(| Healty) = 0.0208 * 0.0208 * 0.0208
= 0.000008998912
Probabilitas = P(Healty) * P(| Healty)
= 0.11 * 0.000008998912
= 0.00000098988032 = **98.988.032 10^{-7}**
7. Kategori Food
P(harga| Food) = 0.0217
P(kentang| Food) = 0.0217
P(impor| Food) = 0.0217
Jadi P(| Food) = 0.0217 * 0.0217 * 0.0217
= 0.000010218313
Probabilitas = P(Food) * P(| Food)
= 0.11 * 0.000010218313
= 0.00000112401443 = **112.401.443 x 10^{-6}**
8. Kategori Travel
P(harga| Travel) = 0.0212
P(kentang| Travel) = 0.0212
P(impor| Travel) = 0.0212
Jadi P(| Travel) = 0.0212 * 0.0212 * 0.0212
= 0.000009528128
Probabilitas = P(Travel)* P(Travel)
= 0.11 * 0.000009528128
= 0.00000104809408 = 104.809.408 x 10^{-6}
9. Kategori Teknologi
P(harga| Teknologi) = 0.0217
P(kentang| Teknologi) = 0.0217
P(impor| Teknologi)= 0.0217
Jadi P(| Teknologi) = 0.0217 * 0.0217 * 0.0217
= 0.000010218313
Probabilitas = P(Teknologi) * P(| Teknologi)
= 0.11 * 0.000010218313
= 0.00000112401443 = **112.401.443 x 10^{-6}**

Jadi kategori dari dokumen artikel berita baru itu termasuk **Kategori Finance** karena memiliki probabilitas paling tinggi yaitu **877.566.976 x 10^{-6}**

3. Simpulan

1. Penelitian ini telah menghasilkan sistem klasifikasi artikel berita terhadap sembilan kategori atau kanal yaitu news, finance, sport, otomotif, entertain, healty, food, travel dan tekno.
2. Sistem klasifikasi artikel berita yang dilakukan dengan menggunakan algoritma *naive bayes classifier* menerima data masukkan berupa data text artikel yang diproses dengan *text mining* yaitu proses *casefolding, tokenizing dan filtering*. Setelah didapat kata kunci dari proses *text mining* akan dilakukan perhitungan dengan *naive bayes classifier* yang menghasilkan keluaran berupa artikel yang sudah terkategori.
3. Ketepatan dalam menentukan kategori pada data artikel baru dipengaruhi oleh data pembelajaran atau data latih pada setiap kategori. Data latih ini berisi kata – kata yang sering muncul pada masing – masing kategori atau kata – kata yang dapat mewakili kategori tertentu.

Daftar Pustaka

[1]. Didin Saepudin. 2013. Implementasi *Text Mining* Untuk Klasifikasi Pengaduan Melalui Web Menggunakan *Naive Bayes Classifier*. Informatika. Cimahi.
[2]. Destuardi & Surya. 2009. Klasifikasi Emosi Untuk Teks Bahasa Indonesia Menggunakan *Naive Bayes Classifier*. Seminar Nasional Pascasarjana IX.
[3]. Amalia, Budi & Antonius. 2008. Sistem Klasifikasi Dan Pencarian Jurnal Dengan Menggunakan Metode *Naive Bayes Classifier* Dan *Vector Space Model*. Informatika.
[4]. L. Dunn, Cherrinton & S. Hollander. 2006. *Enterprise Information System*. Mc Graw-Hill. Singapore.