KLASIFIKASI TEKS DENGAN NAIVE BAYES CLASSIFIER UNTUK PENGELOMPOKAN TEKS ARTIKEL

Nur Anita (1110651094)[1], Bagus Setya Rintyarna S.T,M.Kom [2],
Lutfi Ali Muharom, S.Si [3], Sistem Bisnis Cerdas,
Jurusan Teknik Informatika, Fakultas Teknik,
Universitas Muhammadiyah Jember
E-mail : nuranita80@gmail.com [1],

ABSTRAK

Jumlah teks artikel yang tersedia dalam bentuk digital semakin banyak hampir setiap media massa elektronik memberikan sebuah informasi seperti halnya artikel. Sementara itu, teks dari suatu artikel terkadang memiliki suatu kemiripan antara artikel satu dengan artikel lainnya yang akan membuat pembaca mengalami kesulitan dalam mengklasifikasi. Penelitian ini berusaha untuk mengklasifikasikan beberapa kategori teks artikel dengan mengelompokkan berupa artikel kesehatan dengan menggunakan algoritma *Naive Bayes Classifier*. Klasifikasi ini ditekankan pada teks artikel yang berupa kesehatan, untuk mengetahui nilai akurasi yang akan diukur menggunakan pembobotan dari proses algoritma *Naive Bayes Classifier*. Tahapan dalam metodologi penelitian terdiri dari : pengumpulan dokumen (data set), proses text mining, proses algoritma *Naive Bayes Classifier*, hasil *Naive Bayes Classifier*, dan analisa hasil. Metode *Naive Bayes Classifier* merupakan metode yang digunakan untuk mengklasifikasi. Oleh karena itu untuk menyelesaikan permasalahan ini digunakan metode *Naive Bayes Classifier* sebagai alat untuk mengklasifikasi sebuah teks artikel kesehatan. Hasil pengujian klasifikasi teks artikel kesehatan dengan metode *Naive Bayes Classifier* dapat mengklasifikasi teks artikel kesehatan dengan tingkat keberhasilan *Precision* 91%, *Accuracy* 59%, *Recall* 61% dari nilai rata-rata keseluruhan dokumen percobaan dengan tingkat nilai yang berbeda. Hal ini menunjukkan bahwa metode *Naive Bayes Classifier* tingkat klasifikasi dalam mengelompokkan suatu dokumen belum optimal.

Kata kunci : *Naive Bayes Classifier,* Klasifikasi Teks, Artikel Kesehatan

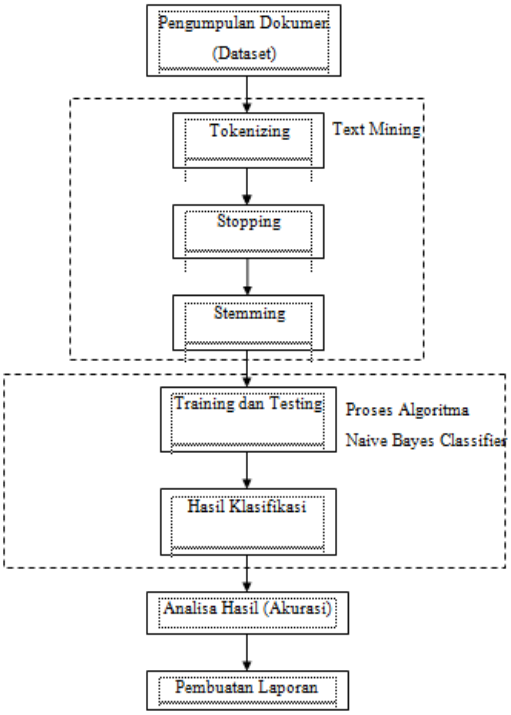
1. PENDAHULUAN

Dalam penentuan kategori artikel khususnya pada artikel kesehatan, manusia terkadang menemukan kesulitan dalam mengklasifikasi atau mengelompokkan suatu artikel berdasarkan masing-masing kategori artikel. Karena teks dari suatu artikel terkadang memiliki suatu kemiripan antara artikel yang satu dengan artikel lainnya yang akan membuat pembaca akan mengalami kesulitan dalam mengklasifikasi. Dalam pengolahan semua kata yang ada pada artikel yang membuat sulit adalah bagaimana mengidentifikasi kata dari artikel tersebut menurut masing-masing kategori yang ada pada artikel kesehatan, diantaranya Spesialis Anak, Spesialis Mata, Spesialis Kulit, Spesialis Jantung, Spesialis Gizi, Spesialis Syaraf, maka digunakanlah metode *Naive Bayes Classifier*. *Naive Bayes Classifier* merupakan salah satu metode machine learning yang

menggunakan perhitungan probabilitas. Konsep dasar yang digunakan oleh *Naive Bayes Classifier* adalah Teorema Bayes, yaitu teorema yang digunakan dalam statistika untuk menghitung suatu peluang, Bayes Optimal Classifier menghitung peluang dari satu kelas dari masing-masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal [Akhmad Basuki. 2006] [1].

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian



Gambar 2.1 Diagram Desain Sistem

Sistem ini akan di implementasikan dengan menggunakan bahasa *PHP: Hypertext Preprocessor* adalah bahasa skrip yang dapat ditanamkan atau disisipkan kedalam HTML. PHP juga banyak dipakai untuk pemrograman situs web dinamis. PHP pertama kali dibuat oleh *Rasmus Lerdorf* pada tahun 1995. [Bunafit Nugroho. 2004] [2].

Dalam pengerjaan Tugas Akhir ini diperlukan langkah-langkah kegiatan penelitian untuk mendapatkan hasil yang maksimal. Untuk itu penulis merencanakan suatu langkah-langkah untuk dapat memaksimalkan dalam pengerjaan Tugas Akhir ini. Langkah-langkah tersebut adalah sebagai berikut :

a. Pengumpulan Dokumen (Dataset)

Pada tahap ini penelitian dilakukan dengan teknik mengumpulkan data-data yang berupa dokumen teks artikel yang ada pada media massa elektronik berbasis website.

b. Text Mining

Setelah proses pengumpulan dokumen selesai kemudian melakukan *scanning* terhadap dokumen teks artikel. Kemudian Memilah dokumen menjadi kalimat. Pemilahan dokumen dilakukan dengan memecah *string* teks dari dokumen yang panjang menjadi kalimat-kalimat mengunakan fungsi *split()*, dengan tanda titik ”.”, tanda tanya ”?” dan tanda seru ”!” sebagai delimiter untuk memotong *string* dokumen. Setelah proses memecah *string* selesai dilanjutkan dengan pemilahan kalimat menjadi perkata *(tokenazing)*. Kemudian menggunakan kata yang tidak penting *(stoplist)* yang pada akhir mengembalikan kedalam bentuk dasar *(stemming)*.

c. Proses Algoritma Naive Bayes Classifier

Proses algoritma Naive Bayes Classifier ini dilakukan setelah melakukan scanning pada dokumen, hasil dari inputan dokumen tesebut kemudian dimasukkan

kedalam perhitungan Naive Bayes Classifier.

d. Hasil Klasifikasi Naive Bayes Classifier

Hasil klasifikasi di dapat pada saat melakukan proses training dan testing terlebih dahulu kemudian di proses dengan algoritma Naive Bayes Classifier, maka setelah melakukan proses tersebut data akan terklasifikasi.

e. Analisa Hasil (Akurasi)

Dapat diketahui klasifikasi atau pengelompokan dokumen teks artikel berdasarkan masing-masing kategori dokumen teks artikel kesehatan.

f. Pembuatan Laporan

Pembuatan laporan ini mencakup semua tahapan yang dilakukan selama penelitian ini. Penulisan dalam laporan ditulis secara jelas agar dapat dimengerti sebagai suatu karya ilmiah, laporan disusun berdasarkan buku pedoman skripsi dan Tugas Akhir Jurusan Teknik Informatika dan Manajemen Informatika Universitas Muhammadiyah Jember.

4. HASIL DAN PEMBAHASAN

Pengujian yang dilakukan pada penelitian ini adalah sebagai berikut :

5 dokumen dataset yang di testing dan 120 dokumen dataset artikel dari 6 kategori artikel , yang diambil dari artikel kesehatan pada suatu media massa elektronik. Di dapakan tabel perhitungan dari nilai *Precision, Recall* dan *Accuracy* seperti berikut [G Salton dan C Buckley. 1990] [3] :

Document	*Accuracy*	*Precision*	*Recall*
Spesialis Anak	60%	91%	61%
Spesialis Mata	55%	91%	58%
Spesialis Kulit	65%	84%	69%
Spesialis Jantung	60%	92%	64%
Spesialis Gizi	55%	100%	55%
Spesialis Syaraf	55%	91%	58%
Average	59%	91%	61%

Spesialis Anak didapatkan *Precision* sebesar 91%, *Accuracy* 60%dan *Recall* 61%, hal ini menunjukkan bahwa sistem dapat mengklasifikasi artikel spesialis Anak dengan baik dan optimal pada sistem karena terdapat kategori artikel yang terklasifikasi dengan benar dan masih ada yang terklasifikasi kedalam kategori lainnya. Pada Spesialis Mata didapatkan *Precision* sebesar 91%, *Accuracy* 55% dan *Recall* 58%, pada hasil *Precision* (ketepatan) tampak baik karena tingkat posterior dari *Precision* terklasifikasi dalam jumlah yang tinggi, sedangkan pada hasil *Accuracy* dan *Recall* dalam jumlah sedikit, hal ini menunjukkan bahwa artikel Spesialis Mata masih ada yang terklasifikasi kedalam kategori artikel lainnya.

Pada Spesialis Kulit didapatkan *Precision* sebesar 84% , *Accuracy* 65% dan *Recall* 69%, hal ini menunjukkan bahwa sistem dapat mengklasifikasi artikel Spesialis Kulit dengan baik pada sistem karena terdapat kategori artikel yang terklasifikasi dengan benar dan masih ada yang terklasifikasi kedalam kategori lainnya. Pada Spesialis Jantung didapatkan *Precision* sebesar 92%, *Accuracy* 60% dan *Recall* 69%, hal ini menunjukkan bahwa sistem dapat mengklasifikasi artikel Spesialis Jantung dengan baik dan optimal pada sistem karena terdapat

kategori artikel yang terklasifikasi dengan benar dan masih ada yang terklasifikasi kedalam kategori lainnya.

Pada Spesialis Gizi didapatkan *Precision* sebesar 100% , *Accuracy* 55% dan *Recall* 55%, pada hasil *Precision* (ketepatan) tampak baik karena tingkat posterior dari *Precision* terklasifikasi dalam jumlah yang tinggi, sedangkan pada hasil *Accuracy* dan *Recall* dalam jumlah sedikit, hal ini menunjukkan bahwa artikel Spesialis Gizi masih ada yang terklasifikasi kedalam kategori artikel lainnya. Pada Spesialis Syaraf didapatkan *Precision* sebesar 91%, *Accuracy* 55% dan *Recall* 58%, pada hasil *Precision* (ketepatan) tampak baik karena tingkat posterior dari *Precision* terklasifikasi dalam jumlah yang tinggi, sedangkan pada hasil *Accuracy* dan *Recall* dalam jumlah sedikit, hal ini menunjukkan bahwa artikel Spesialis Syaraf masih ada yang terklasifikasi kedalam kategori artikel lainnya.

Maka akan didapatkan nilai rata-rata *Precision* sebesar 91%, *Accuracy* 59% dan *Recall* 61% dari total keseluruhan dokumen , hal ini menunjukkan bahwa dari nilai rata-rata tersebut tingkat klasifikasi dari ketepatan, keberhasilan dan akurasi data dalam mengelompokkan suatu dokumen teks artikel belum optimal, karena terdapat kategori artikel yang terklasifikasi dengan benar dan ada kategori artikel yang tidak terklasifikasi pada artikel tersebut.

Accuracy didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual, sedangkan Precision adalah rasio jumlah data yang ditemukan dengan total jumlah data yang telah terklasifikasi kedalam kelasnya masing-masing. Precision mengidentifikasi kualitas himpunan data yang terklasifikasi akan tetapi tidak memandang total data yang relevan. *Recall* ialah perbandingan jumlah dokumen relevan yang didapatkan sistem dengan jumlah seluruh dokumen relevan yang ada dalam koleksi dokumen (terambil ataupun tak terambil system). Suatu sistem atau metode dikatakan optimal apabila nilai precission dan recall tinggi dan nilai precission dan accuracy seimbang.

5. KESIMPULAN DAN SARAN

a. Kesimpulan

1. Metode *Naive Bayes Classifier* ini akan memudahkan pengguna dalam melakukan pengklasifikasian atau pengelompokan artikel.
2. Dokumen teks artikel kesehatan dapat dikelompokkan berdasarkan masing-masing kategori menggunakan klasifikasi teks dengan metode *Naive Bayes Classifier*.
3. Hasil perhitungan dataset dengan metode *Naive Bayes Classifier* didapatkan hasil rata-rata *Precision* sebesar 91%, *Accuracy* 59% dan *Recall* 61% dari total keseluruhan dokumen , hal ini menunjukkan bahwa dari nilai rata-rata tersebut tingkat klasifikasi dari ketepatan, keberhasilan dan akurasi data dalam mengelompokkan suatu dokumen teks artikel belum optimal, karena terdapat kategori artikel yang terklasifikasi dengan benar dan ada kategori artikel yang tidak terklasifikasi pada artikel tersebut.

b. Saran

Penulis ingin memberikan beberapa saran yang mungkin dapat membantu dalam mengembangkan Tugas Akrih ini, saran tersebut adalah :

Dalam penelitian selanjutanya dapat menggunakan data lainnya, karena penggunaan dataset artikel kesehatan belum optimal dalam pengklasifikasian.

REFERENSI

[1] Basuki, Akhmad. 2006.”*Metode Bayes*”. Kuliah PENS-ITS.

[2] Nugroho, Bunafit. 2004. Aplikasi Pemrograman Web Dinamis dengan PHP dan MySQL. Yogyakarta: Gava Media.

[3] [SAL1990] Salton, G. dan Buckley C. 1990. *Improving Retrieval Performance by Relevance Feedback. Cornell University, Ithaca, New York* .