KLASIFIKASI DOKUMEN BERKATEGORI MENGGUNAKAN ALGOTIRMA NAIVE BAYES BERBASIS BERNOULLI

[1]Melisa Ayu Susanti (12 1065 1104), [2]Bagus Setya Ryantiarna, S.Si., M.Si, [3]Daryanto M.kom
Jurusan Teknik Informatika Fakultas Teknik Univertas Muhammadiyah Jember
Email : melisaayus@gmail.com

ABSTRAK

Dalam mengelola informasi dari sekumpulan dokumen dengan jumlah yang besar merupakan sebuah kesulitan untuk mengidentifikasi kata yang ada pada dokumen tersebut menurut masing-masing kategori dari dokumen tersebut diperlukan suatu metode. *Naive Bayes Classifier* merupakan salah satu metode machine learning yang menggunakan perhitungan probabilitas. Klasifikasi teks menggunakan Naive bayes ini ada salah satu model yang dapat membantu kita mengelompokkan dokumen yaitu Bernoulli NB. Penelitian ini berusaha untuk mengklasifikasi kategori dokumen dengan menggunakan algoritma *Naive Bayes* Berbasis *Bernoulli.* Klasifikasi ini ditekankan pada kategori dokumen diantaranya Ekonomi, Kesehatan, Hiburan dan Teknologi, untuk mengetahui nilai akurasi yang akan diukur menggunakan pembobotan proses Algoritma *Naive Bayes Classifier.* Metode Bernoulli NB merupakan metode yang digunakan untuk klasifikasi sebuah teks dari kategori dokumen. Hasil pengujian klasifikasi dokumen kategori dengan menggunakan metode *Naive Bayes* Berbasis *Bernoulli* dapat mengklasifikasikan dokumen kategori dengan tingkat *Presicion* sebesar 70%, *Accuracy* 65% dan *Recall* 70% dari nilai rata-rata keseluruhan dokumen percobaan dengan tingkat nilai berbeda. Hal ini menunjukkan bahwa metode *Naive Bayes* berbasis *Bernoulli* tingkat klasifikasi dalam mengelompokkan suatu dokumen belum optimal.

Kata Kunci : *Naive Bayes, Bernoulli,* Klasifikasi Dokumen Kategori

DOCUMENT CLASSIFICATION USING CATEGORY BASED BERNOULLI NAIVE BAYES ALGORITHM

Melisa Ayu Susanti (12 1065 1104)[1], Bagus Setya Ryantiarna, S.Si. [2], M.Si, Daryanto M.kom[3]

Departement of Informatics, Faculty of Engineering, Muhammadiyah University

Email: melisaayus@gmail.com

ABSTRACT

In managing information from a collection of documents with a large amount of a difficulty to identify words that exist in the document according to each category of the document need a method. Naive Bayes classifier is a machine learning method that uses probability calculations. Text classification using Naive Bayes is no one model that can help us classify documents that Bernoulli NB. This study sought to classify categories of documents using a Naive Bayes algorithm Based Bernoulli. This classification is emphasized in the document categories including Economy, Health, Entertainment and Technology, to determine the accuracy value will be measured using a weighting process Naive Bayes classifier algorithm. NB Bernoulli method is the method used for the classification of a text from the document category. The test results document classification category using the method Based Bernoulli Naive Bayes can classify documents by category Presicion rate of 70%, 65% and Recall Accuracy 70% of the average value of the entire document experiment with different value levels. This shows that a method based on Bernoulli Naive Bayes classification level for classifying a document has not been optimal.

Keyword: Naive Bayes, Bernoulli, Document Classification Category

KLASIFIKASI DOKUMEN BERKATEGORI MENGGUNAKAN ALGOTIRMA NAIVE BAYES BERBASIS BERNOULLI

Meningkatnya perkembangan informasi pada dokumen online dari waktu ke waktu ini menyebabkan meningkatnya pula informasi yang berbentuk teks. Hal ini menimbulkan kesulitan bagi pembaca dalam mencerna sebuah informasi dimana bentuk informasi tersebut dalam format tidak terstruktur. Ketidakteraturan struktur berbentuk teks seperti ini memunculkan penganalisisan teks menjadi lebih terstruktur dengan *text mining. Text mining* mencoba untuk mengekstrasi pola berupa informasi dan pengetahuan yang berguna dari sejumlah besar sumber data melalui identifikasi dan eksplorasi dari suatu pola menarik dengan tujuan mendapatkan informasi yang berguna dari sekumpulan dokumen berbentuk teks yang memiliki format tidak terstruktur.

Dalam klasifikasi teks menggunakan *Naive B*ayes *classifier* ini ada salah satu model yang dapat membantu kita mengelompokkan dokumen yaitu *Bernoulli NB. Bernoulli* merupakan fungsi *Naive Bayes Classifier* yang menggunakan unsur biner mengambil nilai 1 bila kata yang sesuai terdapat dalam dokumen dan 0 bila kata tersebut tidak ada.

Dalam penelitian ini menjelaskan tentang metode yang akan digunakan selama penelitian berlangsung. Penelitian ini melalui beberapa tahap diantaranya Studi Literatur, Perancangan Sistem dan Pengujian. Pengujian difokuskan pada pengujian keakuratan klasifikasi dokumen. Pengujian kinerja *Naive Bayes* akan di ukur dengan parameter uji *recall* dan *precision.*

Mengenai hasil penelitian yang dilakukan penulis terhadap akurasi dari klasifikasi *naive bayes* dengan objek penelitian sebanyak 60 dokumen. Pada sistem ini hanya terdapat satu aktor yaitu *user*. Ketika pertama kali menjalankan sistem, *user* diharuskan melakukan pembelajaran sistem terlebih dahulu sesuai dengan kebutuhan sistem, *user* dapat menginputkan dokumen yang telah diketahui kategorinya dan melakukan proses pembelajaran. Selanjutnya *user* dapat melakukan klasifikasi dokumen dengan menginputkan dokumen yang belum diketahui kategorinya dan sistem mengklasifikasikan secara otomatis berdasarkan kategori yang ada. *User interface* pada klasifikasi *naive bayes* ini terdiri dari

beberapa form diantaranya form kategori, form filtering, form dokumen pembelajaran dan form dokumen klasifikasi.

Pada pengujian untuk mengetahui nilai *recall* dan *precision* yang dilakukan pada sistem klasifikasi ini, dengan menguji data testing pada beberapa kali proses training. Hasil Pengujian yang diperoleh pada pengujian ini akan dijelaskan dalam Tabel 4.4 dibawah ini:

Tabel 4.7 Nilai *Accuracy*, *Precision* dan *Recall*

Percobaan ke-	*Accuracy*	*Precision*	*Recall*
1	53%	57%	17%
2	59%	75%	26%
3	55%	53%	85%
4	55%	52%	58%
5	64%	80%	50%
Rata-rata	57%	63%	47%

Dari hasil pengujian Naive Bayes Classifier dapat diketahui kategori dokumen yang benar dan dokumen yang terklasifikasi pada kategori dokumen tersebut, serta diketahui berapa presentase tingkat keakurasian algoritma *Naive Bayes Classifier*.

Dari hasil pengujian pada tabel 4.7 ketahui nilai *precision, recall* dan *accuracy* untuk setiap percobaan. Perhitungan rata-rata dari semua percobaan *Precision* 63%, *Accuracy* 57% dan *Recall* 47%. Nilai *precision* tertinggi dari semua percobaan yaitu 80% sedangkan nilai precision terendah dari semua percobaan yaitu 52%. Untuk nilai recall tertinggi dari semua percobaan yaitu 85% sedangkan nilai recall terendah dari semua percobaan yaitu 17%. Untuk nilai accuracy tertinggi dari semua percobaan yaitu 64% sedangkan nilai recall terendah dari semua percobaan yaitu 53% Hal ini menunjukkan tingkat klasifikasi dari ketepatan, keberhasilan dan akurasi data dalam suatu kategori dokumen teks belum optimal, karena terdapat kategori dokumen yang terklasifikasi dengan benar dan ada kategori dokumen yang tidak terklasifikasi pada kategori tersebut.

Kesimpulan

Berdasarkan analisis dan pengujian yang dilakukan pada bab sebelumnya, maka kesimpulan yang dapat diambil adalah sebagai berikut :

1. Pengujian pada dokumen percobaan ke-5 dengan menggunakan 60% data testing menghasilkan nilai *accuracy* 64%, *precision* 80% dan *recall* 50%.. Dari semua percobaan tingkat klasifikasi dari ketepatan, keberhasilan dan akurasi data dalam percobaan ke-5 menghasilkan nilai *accuracy* yang tinggi, nilai precision yang tinggi dan nilai recall yang relatif tinggi. Hal ini menunjukkan semakin besar presentase data testing maka hasil pengujian semakin baik.
2. Penentuan data training dapat mempengaruhi hasil pengujian, karena pola data training tersebut dijadikan sebagai rule untuk menentukan kelas pada data testing. Sehingga besar atau kecilnya presentase tingkat *precision*, *recall*, dan *accuracy* dipengaruhi juga oleh penentuan data training.

Saran

Adapun beberapa saran dari penulis untuk pengembangan tesis ini adalah :

1. Data yang digunakan lebih baik dalam jumlah yang besar untuk data pembelajaran dengan semakin banyaknya fitur kata sehingga persentasi akurasi lebih tinggi.
2. Penelitian ini dapat dilanjutkan dengan menggunakan kategori dan sub kategori lainnya yang lebih bervariasi lagi sehingga klasifikasi dokumen lebih tepat.

DAFTAR PUSTAKA

Abidin, Taufik fuadi. 2009. *Bayesian Teorem*, Data Mining dan Information Retrival Research Group

Artikel Pilihan. 2008. Artikel Berita. Diambil dari : http://www.kompasiana.com

(16 November 2015).

Basuki, Akhmad. 2006."Metode Bayes" Kuliah PENS-ITS

Information Retrieval. 2011. Pengertian Information Retrival. Diambil dari : https://arevoblogs.files.wordpress.com (3 November 2015).

Hiroshi Shimodaira. 2014. "Text Classification Using Naive Bayes".Informatics 2B

Muthafa, Aziz. 2009. *Klasifikasi otomatis Dokumen Berita Kejadian Berbahasa Indonesia.* Malang : Fakultas Sains dan Teknologi. Universitas Islam Negeri Maulana Malik Ibrahim

Putri, Arini. 2013. Klasifikasi Dokumen Teks Menggunakan Metode *Support Vector Machine* dengan pemilihan vitur *Chi-square*. Bogor : Institut Pertanian Bogor.