EDUKOM 2 (1) (2015)

Edu Komputika Journal

http://journal.unnes.ac.id/sju/index.php/edukom

PEMILIHAN *FEATURE* DENGAN *CHI SQUARE* DALAM ALGORITMA *NAÏVE BAYES* UNTUK KLASIFIKASI BERITA

Alfian Nur Rahmad✉ **dan Feddy Setio Pribadi**

Jurusan Teknik Elektro, Fakultas Teknik, Universitas Negeri Semarang, Indonesia

Info Artikel

Sejarah Artikel:
Diterima April 2015
Disetujui Mei 2015
Dipublikasikan Juni 2015

Keywords:
Chi Square, classification, feature selection, Naïve Bayes, news

Abstrak

Klasifikasi berita secara manual tidak mungkin dilakukan. Klasifikasi otomatis banyak dilakukan dengan algoritma naïve bayes, tetapi jumlah feature kata yang banyak dapat mengurangi akurasi klasifikasi. Penelitian ini bertujuan untuk mengetahui penerapan, pengaruh dan nilai recall, precision, f-measure dan akurasi dari pemilihan feature Chi Square terhadap kinerja algoritma Naïve Bayes untuk mengklasifikasikan teks berita secara otomatis. Pada penelitian ini diterapkan teknik pemilihan feature dengan Chi Square dalam Algoritma Naïve Bayes. Data penelitian diambil dari www.kompas.com sebanyak 1350 buah sebagai data latih dan 150 buah sebagai data uji. Pengujian dilakukan dengan mengklasifikasikan berita tanpa pemilihan feature Chi Square dan mengklasifikasikan berita dengan menerapkan pemilihan feature Chi Square dengan taraf nyata α 0.05, 0.01, 0.005, dan 0.001. Selanjutnya akan dievaluasi dengan metode evaluasi recall, precision, f-measure dan akurasi. Dari klasifikasi berita otomatis tanpa pemilihan feature yang dilakukan, diperoleh hasil recall 96.67%, precision 96.75%, f-measure 96.68% dan akurasi 96.67%. Sedangkan klasifikasi berita dengan pemilihan feature menggunakan chi square pada taraf nyata α 0.05, 0.01, 0.005, dan 0.001 diperoleh hasil yang sama yaitu recall 98%, precision 98%, f-measure 97.99%, dan akurasi 98%. Dari hasil tersebut, dapat diketahui bahwa pemilihan feature menggunakan chi square dapat mempengaruhi kinerja algoritma Naïve Bayes untuk mengklasifikasikan berita secara otomatis.

Abstract

Classification of news manually impossible. Automatic classification lot to do with the naïve Bayes algorithm, but the number of words that many features can reduce the accuracy of the classification. This study aims to determine the application, influence and value of recall, precision, f-measure and accuracy of election Chi Square feature of the performance Naïve Bayes algorithm to automatically classify news text. In this study feature selection techniques applied by Chi Square in Naïve Bayes algorithm. Data were taken from as many as 1350 pieces www.kompas.com as training data and 150 as test data. Testing is done by classifying feature election news without Chi Square and classifying news by applying the Chi Square feature selection with significance level α 0:05, 0:01, 0.005, and 0.001. Next will be evaluated by the evaluation method of recall, precision, f-measure and accuracy. Automatic classification of news without selecting a feature that is done, the result recall 96.67%, 96.75% precision, f-measure 96.68% and 96.67% accuracy. While the classification of news with feature selection using the chi square on the real level α 0:05, 0:01, 0005, and 0001 obtained the same result, namely 98% recall, 98% precision, f-measure 97.99%, and accuracy 98%. From these results, it is known that the selection of the feature using the chi square can affect the performance Naïve Bayes algorithm to automatically classify news.

✉ Alamat korespondensi:
Gedung E6 Lantai 2 FT Unnes
Kampus Sekaran, Gunungpati, Semarang, 50229
E-mail: alfian.ptik@gmail.com

ISSN 2252-6811

PENDAHULUAN

Berita saat ini menjadi sangat penting bagi masyarakat. Setiap hari orang mencari dan membaca berita untuk mendapatkan suatu informasi (Liliana, Hardianto, dan Ridok, 2011). Salah satu media penyebaran berita yang banyak diminati pada zaman ini adalah media online (Musthafa, 2009). Beberapa situs berita memiliki kategori seperti berita politik, olah raga, ekonomi, kesehatan, dan lain-lain. Dalam membagi berita kedalam kategori-kategori tersebut untuk saat ini masih dilakukan secara manual oleh manusia, sehingga membutuhkan biaya yang mahal (Kompan dan Beilikova, 2011). Selain itu pengklasifikasian secara manual membutuhkan waktu yang lebih lama (Nindhi dan Gupta, 2012). Oleh karena itu dibutuhkan klasifikasi berita secara otomatis agar dapat mengurangi biaya dan waktu.

Penyelesaian masalah pengklasifikasian suatu berita yang sebagian besar berbentuk teks dapat dilakukan dengan menggunakan metode text minning. Text mining merupakan variasi dari data mining yang berusaha menemukan pola yang menarik dari sekumpulan data tekstual yeng berjumlah besar (Feldman & Sanger, 2007). Salah satu tipe dalam text mining yaitu klasifikasi dokumen. Klasifikasi dokumen adalah suatu proses pengelompokan kumpulan dokumen ke dalam kelasnya berdasarkan kontennya (Odeh et al, 2014).

Proses klasifikasi teks sebelumnya telah banyak dilakukan dengan berbagai metode, diantaranya Naïve Bayes (Ting, Ip, dan Tsang, 2011), self organizing maps (Isa, Kallimani and Lee, 2009), K-Nearest Neighbor (Trstenjak, Mikac, dan Donko, 2014), dan Decision Trees (Saad, 2010). S.L Ting, Ip, dan Tsang (2011) melakukan perbandingan empat metode klasifikasi teks. Hasilnya tingkat akurasi metode Naïve Bayes mencapai 97% , support vector machines 96,9%, neural network 93 % dan Decision Trees 91,1%. Penelitian tentang klasifikasi berita berbahasa Indonesia dengan Naïve Bayes Classifier (NBC) sebelumnya sudah pernah dilakukan oleh Yudi Wibisono (2008). NBC terbukti dapat digunakan secara efektif untuk mengklasifikasikan berita secara otomatis dengan akurasi mencapai 90.23%. Penelitian lain tentang klasifikasi teks berita dengan NBC dilakukan oleh Amir Hamzah (2012) menunjukkan hasil akurasi 91%.

Berdasarkan hasil uji coba awal yang dilakukan, diketahui bahwa klasifikasi teks memiliki masalah yang berkaitan dengan besarnya dimensi data. Hal ini dikarenakan pada klasifikasi teks, suatu dokumen direpresentasikan sebagai kumpulan dari kata-kata (*bag of word*), dimana tiap-tiap kata dalam dokumen tersebut tidak bergantung satu sama lain (Schneider, 2005). Kata atau *term* adalah atribut/*feature* dalam menentukan kelas dari sebuah dokumen atau teks (Hamzah 2012). Sehingga diperlukan pemilihan kata atau *feature* yang memiliki pengaruh besar atau merupakan ciri dari suatu kelas. Dalam penelitian sebelumnya yang dilakukan Hamzah (2012), pemilihan feature kata menggunakan jumlah kemunculam kata pada dokumen mendapatkan hasil bahwa kata yang muncul minimal pada 4 atau 5 dokumen menghasilkan akurasi 91%. Namun masih diperlukan penelitian dengan teknik yang lebih baik dalam melakukan pemilihan feature kata yang digunakan sebagai dasar klasifikasi untuk mencapai hasil yang optimal (Hamzah, 2012).

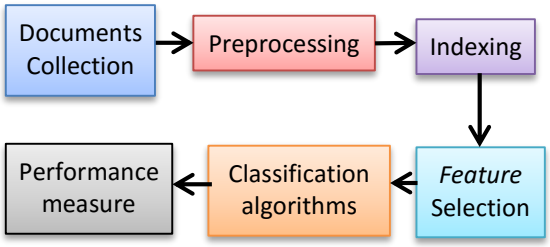
Beberapa teknik yang digunakan untuk melakukan pemilihan feature dokumen antara lain Document Frequency Thresholding (DF), Information Gain (IG), Mutual Information (MI), Term strength (TS) dan Chi-square testing (X2) (Sun, 2009). Dalam pengklasifikasian sebuah dokumen, Chi Square adalah salah satu supervised feature selection yang mampu menghilangkan banyak feature tanpa mengurangi tingkat akurasi (Sun, Wang, dan Xu, 2009). Hasil penelitian Alshalabi et al (2013) menunjukkan bahwa penggabungan antara metode klasifikasi Naïve Bayes dan pemilihan feature Chi Square dalam klasifikasi teks menghasilkan performa yang sangat baik (Alshalabi et al, 2013).

Berdasarkan latar belakang di atas, maka dilakukan penelitian dengan judul "Pemilihan Feature dengan Chi Square dalam Algoritma Naïve Bayes untuk Klasifikasi Berita". Pemilihan

feature menggunakan Chi Square pada algoritma Naïve Bayes untuk mengklasifikasikan teks berita, diharapkan dapat meningkatkan akurasi dan performa dari klasifikasi teks berita otomatis.

METODE PENELITIAN

Penelitian ini dilaksanakan dalam beberapa tahapan sesuai dengan proses klasifikasi teks seperti pada gambar 1.



Gambar 1 Proses Klasifikasi Teks (Bhumika, Sehra, dan Nayyar, 2013)

Documents Collection

Pada tahap awal yaitu documents collection atau pengumpulan dokumen. Pada tahap ini dilakukan pengumpulan dokumen artikel berita dari portal berita www.kompas.com dari bulan Maret 2015 sampai dengan Juni 2015. Berita di*copy* lalu dimasukkan kedalam *notepad* dan disimpan dengan format *.txt. Kategori berita yang diambil yaitu berita ekonomi, olahraga, teknologi, kesehatan, dan hiburan dengan jumlah berita tiap kategori sebanyak 300 dokumen. Data dibagi menjadi dua bagian yaitu data pembelajaran dan data pengujian. Dari jumlah seluruh dokumen, jumlah data pembelajaran yang digunakan sebanyak 90% dan jumlah data pengujian sebanyak 10%.

Preprocessing

Pada tahap preprocessing dilakukan penyeragaman bentuk kata, penghilangan noise, dan mengurangi volume kosakata dari suatu dokumen berita. Tahap ini meliputi tokenizing dan filtering.

1. Tokenizing

Pada tahap ini dokumen berita diseragamkan bentuk hurufnya menjadi huruf kecil (case folding). Setelah itu karakter yang bukan termasuk huruf dibuang. Selanjutnya dokumen dipotong menjadi kata per kata.

2. Filtering

Pada tahap filtering hasil kata dari proses tokenizing disaring dengan menggunakan stopwords, sehingga diperoleh kata-kata penting untuk proses selanjutnya.

Indexing

Pada tahap indexing dilakukan perhitungan jumlah kemunculan tiap kata dalam tiap dokumen berita untuk tiap kategori. Selain itu dilakukan juga perhitungan jumlah dokumen yang terdapat kata tersebut. Sehingga diperoleh nilai kemunculan kata (term frekuensi) dan nilai kemunculan dokumen (documents frekuensi) dari setiap kata untuk tiap kategori yang selanjutnya akan disimpan.

Feature Selection

Tahap selanjutnya pada klasifikasi teks yaitu *feature selection.* Secara umum, dasar ide algoritma *feature selection* yaitu mencari semua kemungkinan kombinasi dari atribut dalam data untuk menemukan subset dari feature yang terbaik untuk prediksi (Ting, Ip, dan Tsang, 2011). *Feature selection* atau selanjutnya disebut pemilihan *feature* dapat digunakan untuk meningkatkan kinerja dari klasifikasi teks dalam hal kecepatan pembelajaran dan efektifitas (Alshalabi et al, 2013).

Dalam penelitian ini digunakan pemilihan *feature* dengan *chi square* (x^2). Pada tahap ini, tiap kata yang diperoleh dihitung menggunakan persamaan sebagai berikut :

$$\chi^2(t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad [1]$$

Keterangan :

A = banyaknya dokumen dalam kategori c yang mengandung term t

B = banyaknya dokumen yang bukan kategori c tetapi mengandung term t

C = banyaknya dokumen dalam kategori c tetapi tidak mengandung term t

D = banyaknya dokumen yang bukan kategori c dan tidak mengandung term t

N = total keseluruhan dokumen

Kemudian diambil kata-kata yang memiliki nilai chi square lebih besar dari nilai kritis pada taraf nyata α 0.05, 0.01, 0.005, dan 0.001 yaitu 3.84, 6.63, 7.83, dan 10.83. Semakin kecil taraf nyata α dan semakin besar nilai kritis, maka jumlah feature yang dihasilkan akan semakin sedikit dan akan mengurangi beban database. Sehingga menghasilkan kata-kata yang merupakan ciri dari suatu kategori tertentu yang selanjutnya akan digunakan dalam proses klasifikasi.

Classification Algorithms

Tahap ini merupakan tahap utama dari proses klasifikasi berita. Algoritma yang digunakan dalam penelitian ini yaitu algoritma Naïve Bayes. Naïve Bayes merupakan salah satu metode supervised document classification. Metode ini sering digunakan dalam menyelesaikan masalah dalam bidang mesin pembelajaran karena metode ini dikenal memiliki tingkat akurasi yang tinggi dengan perhitungan sederhana (Anggarwal dan Zhai, 2012 : 176).

Dengan aturan Bayes maka untuk klasifikasi dokumen dapat dinyatakan bahwa (Schneider, 2005) :

$$p(c_j \mid d) = \frac{p(c_j)p(d \mid c_j)}{p(d)} \quad [2]$$

Dimana cj adalah kategori dokumen yang akan diklasifikasikan, dan p(cj) merupakan probabilitas dari kategori teks cj. Pada saat proses pengklasifikasian dokumen teks, maka pendekatan Bayes akan memilih kategori yang memiliki probabilitas paling tinggi (CMAP) yaitu :

$$C_{MAP} = \text{argmax} \frac{p(cj)p(d \mid cj)}{p(d)} \quad [3]$$

Nilai p(d) dapat diabaikan karena nilainya adalah konstan untuk semua cj, sehingga persamaan [3] dapat dituliskan sebagai berikut:

$$C_{MAP} = \text{argmax}\, p(c_j)\, p(d \mid c_j) \quad [4]$$

Probabilitas p(cj) dapat diestimasi dengan menghitung jumlah dokumen training pada setiap kategori cj. Sedangkan untuk menghitung distribusi dokumen pada tiap kelas, p(d | cj) tidak dapat secara langsung ditentukan. Diasumsikan bahwa sebuah dokumen terdiri dari unit yang lebih kecil, biasanya merupakan kumpulan dari kata (w1,w2,w3,...)(Schneider, 2005). Dengan pendekatan Naïve Bayes yang mengasumsikan bahwa setiap kata tidak bergantung satu sama lain, maka perhitungan dapat lebih disederhanakan dan dapat dituliskan sebagai berikut:

$$p(d \mid c_j) = \prod_{i=1}^{n} p(w_i \mid c_j) \quad [5]$$

Dengan menggunakan persamaan [5], maka persamaan [4] dapat dituliskan menjadi :

$$C_{MAP} = \text{argmax}\ p(c_j) \prod_{i=1}^{n} p(w_i \mid c_j) \quad [6]$$

Nilai p(cj) dan p(wi | cj) dihitung pada saat proses pembelajaran dimana persamaannya adalah sebagai berikut :

$$p(c_j) = \frac{|\text{docs j}|}{|\text{contoh}|} \quad [7]$$

$$p(w_i \mid c_j) = \frac{1+n_i}{|C|+n(\text{kosakata})} \quad [8]$$

p(wi | cj) = probabilitas kata wi pada kategori cj

|docs j| = jumlah dokumen pada kategori j

|contoh| = jumlah seluruh dokumen sampel yang digunakan dalam proses training

ni= frekuensi kemunculan kata wi pada kategori cj

| C | = jumlah semua kata pada kategori cj

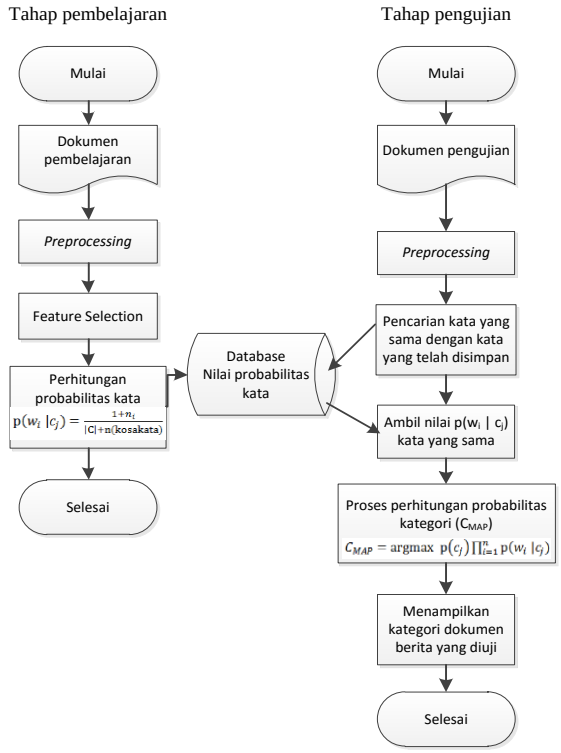
n(kosakata) = jumlah kata yang unik pada semua data training.

Perkalian nilai desimal yang terlalu besar akan membuat nilai terlalu banyak angka di

belakang koma (,) (floating point underflow) sehingga data tidak dapat ditampung dalam komputer. Untuk permasalahan ini dapat digunakan penambahan logaritma dalam perhitungan, dimana log(xy) = log(x) + log(y), sehingga persamaan [6] dirubah menjadi sebagai berikut (Manning, Raghavan, dan Schutze, 2009) :

$$C_{MAP} = \text{argmax}\ \log \mathrm{p}(c_j) + \Sigma_{1 \le i \le n} \log \mathrm{p}(w_i|c_j)$$ [9]

Dalam klasifikasi berita menggunakan algoritma Naïve Bayes terdapat dua tahapan utama yaitu pembelajaran dan pengujian. Alur sistem klasifikasi berita secara otomatis seperti gambar 2.



Gambar 2 Diagram Alir Arsitektur Sistem Klasifikasi Berita Otomatis

a. Performance Measure

Tahap terakhir adalah penghitungan performance measure, yaitu mengevaluasi hasil percobaan dan menganalisis kinerja klasifikasi teks dengan menghitung nilai recall, precision, f-measure dan akurasi.

Pengujian klasifikasi berita secara otomatis dilakukan dengan dua tahap, yaitu pengklasifikasian berita tanpa menerapkan pemilihan *feature Chi Square* dan pengklasifikasian berita dengan menerapkan pemilihan *feature Chi Square* pada taraf nyata 0.05,0.01,0.005, dan 0.001.

HASIL DAN PEMBAHASAN

HASIL PENGUJIAN

Data pengujian sebanyak 150 dokumen yang terbagi kedalam 5 kategori berita yaitu teknologi, ekonomi, olahraga, kesehatan, dan hiburan dengan tiap kategori berjumlah sama yaitu 30 dokumen, diperoleh hasil klasifikasi terhadap pemilihan feature dan tanpa pemilihan feature seperti berikut :

1. Hasil Klasifikasi Tanpa Pemilihan Feature

Hasil klasifikasi tanpa pemilihan feature yang dilakukan, terlihat seperti pada gambar 3. Hasil yang diperoleh dari pengujian berita dengan kategori berita teknologi yaitu 30 dokumen yang sesuai dari 30 dokumen yang diujikan, untuk kategori berita ekonomi diperoleh 28 dokumen sesuai kategori dan 2 dokumen tidak sesuai kategori. Kategori berita olahraga, kesehatan dan hiburan memiliki hasil yang sama yakni 29 dokumen sesuai kategori dan 1 dokumen tidak sesuai kategori.

Tabel Confusion Matrix

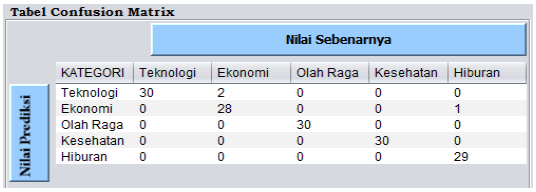
		Nilai Sebenarnya				
	KATEGORI	Teknologi	Ekonomi	Olah Raga	Kesehatan	Hiburan
Nilai Prediksi	Teknologi	30	2	0	0	0
	Ekonomi	0	28	0	1	1
	Olah Raga	0	0	29	0	0
	Kesehatan	0	0	0	29	0
	Hiburan	0	0	1	0	29

Gambar 3 Hasil Klasifikasi Tanpa Pemilihan Feature

1. Hasil Klasifikasi dengan Pemilihan Feature

Hasil klasifikasi dengan pemilihan feature yang telah dilakukan, terlihat seperti pada gambar 3. Pemilihan feature dilakukan dengan menggunakan taraf nyata α 0.05, 0.01, 0.005, dan 0.001. Hasil klasifikasi berita dari empat taraf

nyata tersebut, diperoleh hasil yang sama yaitu dari 30 dokumen yang diujikan, berita dengan kategori berita teknologi, olahraga dan kesehatan menghasilkan 30 dokumen yang sesuai kategori, untuk kategori berita ekonomi diperoleh 28 dokumen sesuai kategori dan 2 dokumen tidak sesuai kategori, serta untuk kategori hiburan diperoleh hasil 29 dokumen sesuai kategori dan 1 dokumen tidak sesuai kategori.


Tabel Confusion Matrix

		Nilai Sebenarnya				
	KATEGORI	Teknologi	Ekonomi	Olah Raga	Kesehatan	Hiburan
Nilai Prediksi	Teknologi	30	2	0	0	0
	Ekonomi	0	28	0	0	1
	Olah Raga	0	0	30	0	0
	Kesehatan	0	0	0	30	0
	Hiburan	0	0	0	0	29

Gambar 4 Hasil Klasifikasi Setelah Dilakukan Pemilihan Feature dengan Taraf Nyata α 0.05, 0.01, 0.005, dan 0.001

Hasil klasifikasi yang telah diperoleh, selanjutnya dihitung nilai *Precision, Recall, F-measure* dan akurasinya. Hasil perhitungan yang dilakukan yaitu seperti berikut:

1. Hasil Evaluasi Klasifikasi Tanpa Pemilihan Feature

Hasil evaluasi klasifikasi tanpa pemilihan feature yang diperoleh yaitu seperti pada tabel 1.

Tabel 1 Hasil Evaluasi Klasifikasi Tanpa Dilakukan Pemilihan Feature

Evaluasi	**Rata-rata**
Recall	96.667%
Precision	96.75%
F-measure	96.677%
Akurasi	96.667%

2. Hasil Evaluasi Klasifikasi dengan Pemilihan Feature

Hasil evaluasi klasifikasi tanpa pemilihan feature yang diperoleh yaitu seperti pada tabel 2.

Tabel 2 Hasil Evaluasi Klasifikasi dengan Pemilihan Feature Chi Square Pada Taraf Nyata α 0.05, 0.01, 0.005, dan 0.001

Evaluasi	Rata-rata
Recall	98%
Precision	98.06%
F-measure	97.999%
Akurasi	98%

Pada tabel 2 terlihat bahwa penggunaan taraf nyata yang berbeda tidak mempengaruhi hasil klasifikasi berita. Namun, penggunaan taraf nyata tersebut berpengaruh terhadap jumlah feature yang digunakan dalam pengklasifikasian. Jumlah feature yang dihasilkan pada tahap pembelajaran sesuai dengan uji coba yang dilakukan terlihat seperti pada tabel 3.

Tabel 3 Jumlah Feature Pada Beberapa Uji Coba

α \ kategori	Total
tanpa X^2	39.172
0.05	22.342
0.01	8.371
0.005	7.724
0.001	4.820

Perbandingan hasil evaluasi terhadap sistem dari beberapa uji coba yang dilakukan terlihat seperti pada tabel 4.

Tabel 4. Perbandingan Recall, Precision, F-measure dan Akurasi Pada Beberapa Uji Coba

α	Recall	Precision	F-measure	Akurasi
tanpa X^2	96.67%	96.75%	96.68%	96.67%
0.05	98%	98.06%	97.99%	98%
0.01	98%	98.06%	97.99%	98%
0.005	98%	98.06%	97.99%	98%
0.001	98%	98.06%	97.99%	98%

PEMBAHASAN

Pada penelitian ini klasifikasi berita dilakukan dengan beberapa uji coba antara lain pengklasifikasian tanpa dilakukan pemilihan feature, klasifikasi dengan pemilihan feature menggunakan Chi Square pada taraf nyata α yang berbeda yaitu 0.05, 0.01, 0.005, dan 0.001. Pengklasifikasian dilakukan dengan beberapa uji

coba yang berbeda dimaksudkan untuk mengetahui pengaruh pemilihan feature terhadap hasil dari klasifikasi.

Berdasarkan penelitian yang telah dilakukan, pengklasifikasian berita secara otomatis menggunakan algoritma Naïve Bayes tanpa pemilihan feature diperoleh hasil Recall, Precision, F-measure sebesar 96.67%, 96.75 % dan 96.68 %. Sedangkan pengklasifikasian berita secara otomatis menggunakan algoritma Naïve Bayes dengan pemilihan feature Chi Square pada taraf nyata 0.05, 0.01, 0.005, dan 0.001 diperoleh hasil Recall, Precision, F-measure yang sama yaitu 98%, 98.06% dan 97.99%. Hal ini membuktikan bahwa penggunaan Chi Square untuk pemilihan feature dalam proses klasifikasi berita otomatis dapat meningkatkan hasil baik Recall, Precision, F-measure, dan akurasi meskipun peningkatannya tidak terlalu signifikan dan penggunaan taraf nyata α yang berbeda tidak mempengaruhi hasil klasifikasi. Tetapi, penggunaan taraf nyata α sangat berpengaruh terhadap jumlah feature yang didapat. Dengan menggunakan jumlah feature yang sedikit yaitu 4.820 pada taraf nyata α 0.001 memiliki hasil klasifikasi yang sama dengan jumlah feature 22.342 pada taraf nyata α 0.05.

Sejalan dengan penelitian Yoga Herawan (2011), penggunaan Chi Square dalam pemilihan feature dapat meningkatkan hasil klasifikasi. Akurasi klasifikasi sistem yang didapat dari hasil penelitian pengklasifikasian berita secara otomatis menggunakan algoritma Naïve Bayes tanpa pemilihan feature adalah 96.67% dan dengan menggunakan pemilihan feature baik pada taraf nyata α 0.05, 0.01, 0.005, dan 0.001 adalah 98%. Hasil yang diperoleh dari penggunaan taraf nyata α tersebut, berbeda dengan hasil penelitian Yoga Herawan (2011) yang mampu membuktikan penggunaan taraf nyata α 0.001 lebih baik dibandingkan dengan 0.01 untuk meningkatkan kinerja klasifikasi.

Penggunaan Chi square untuk pemilihan feature dalam pengklasifikasian berita secara otomatis menggunakan algoritma Naïve Bayes yang dilakukan, menghasilkan akurasi klasifikasi yang lebih baik dibandingkan dengan pemilihan feature kata pada penelitian Amir Hamzah (2012) yang menggunakan filter minimal kata muncul dalam 4 dokumen. Pada penelitian Hamzah tersebut, diperoleh akurasi sebesar 91% sedangkan dalam penelitian ini 98%.

KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan bahwa:

1. Pemilihan feature dengan menggunakan Chi Square dalam algoritma Naïve Bayes untuk klasifikasi berita otomatis diterapkan dengan memakai taraf nyata α 0.05, 0.01, 0.005, dan 0.001.
2. Pemilihan feature dengan menggunakan Chi Square dalam algoritma Naïve Bayes untuk klasfikasi berita otomatis, berpengaruh terhadap pengurangan jumlah feature yang diperoleh. Pada taraf nyata α 0.05, jumlah kata yang diperoleh yaitu 22.342 dari jumlah kata awal yaitu 39.179, berkurang sebanyak 42.96 %. Untuk taraf nyata α 0.01 berkurang 78.63%, taraf nyata α 0.005 berkurang 80.28%, taraf nyata α 0.001 berkurang 87.70%. Berkurangnya jumlah feature dapat mengurangi beban data dalam database.
3. Nilai recall yang diperoleh dari klasifikasi berita otomatis tanpa pemilihan feature yaitu 96.67% dan setelah menggunakan pemilihan feature menjadi 98%, Precision dari 96.75% menjadi 98%.06, F-measure dari 96.68% menjadi 97.99% dan akurasi dari 96.67% menjadi 98%. Berdasarkan hasil tersebut, dapat diketahui bahwa pemilihan feature Chi Square mempengaruhi kinerja algoritma Naïve Bayes dan dapat meningkatkan hasil klasifikasi.

SARAN

Berdasarkan hasil penelitian yang dilakukan, berikut beberapa saran untuk penelitian selanjutnya:

1. Perlu adanya penelitian menggunakan teknik pemilihan feature yang lain untuk mengklasifikasikan berita berbahasa indonesia.
2. Perlu adanya penelitian yang lebih mendalam untuk mengetahui pengaruh jumlah dokumen terhadap kinerja algoritma naïve bayes.
3. Perlu dilakukan penelitian lebih lanjut mengenai klasifikasi berita otomatis dengan algoritma Naïve Bayes menggunakan data pelatihan dan pengujian yang memiliki rentang waktu berbeda.

UCAPAN TERIMA KASIH

Ucapan terimakasih ditujukan kepada Prof. Dr. Fathur Rokhman, M.Hum., Drs. H, Muhammad Harlanu, M.Pd., Drs. Suryono, M.T., Feddy Setio Pribadi, S.Pd., M.T., serta seluruh dosen Program Studi Pendidikan Teknik Informatika dan Komputer Jurusan Teknik Elektro Fakultas Teknik Unnes.

DAFTAR PUSTAKA

Alshalabi, Hamood, Sabrina Tiun, Nazlia Omar, dan Mohammed Albared. 2013. Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization. ICEEI 2013. Universiti Kebangsaan Malaysia. Malaysia.

Feldman, Ronen dan James Sanger. 2007. The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press. New York.

Hamzah, Amir. 2012. Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis. Prosiding Seminar.

Herawan, Yoga. 2011. Ekstraksi Ciri Dokumen Tumbuhan Obat Menggunakan Chi-Kuadrat dengan Klasifikasi Naïve Bayes. IPB. Bogor.

Isa, Dino, V.P. Kallimani, dan Lam Hong Lee. 2009. Using The Self Organizing Map For Clustering Of Text Documents. Expert Systems with Applications 36 (2009) 9584–9591. Elsevier.

Kompan, Michal dan Maria Beilikova. 2011. News Article Classification Based on a Vector Representation Including Words' Collocations. Third International Conference on Software, Services and Semantic Technologies S3T 2011 Advances in Intelligent and Soft Computing Volume 101, 2011, pp 1-8. Springer Berlin Heidelberg.

Liliana, Dewi Y., Agung Hardianto, dan M. Ridok. 2011. Indonesian News Classification using Support Vector Machine. World Academy of Science, Engineering and Technology Vol:5 2011-09-21.

Musthafa, Aziz. 2009. Klasifikasi Otomatis Dokumen Berita Kejadian Berbahasa Indonesia. Skripsi. Jurusan Teknik Informatika, Fakultas Sains dan Teknologi, Universitas Islam Negeri (UIN) Maulana Malik Ibrahim. Malang.

Nindhi dan Vispal Gupta. 2012. Punjabi Text Classification using Naïve Bayes, Centroid and Hybrid Approach. http://airccj.org/CSCP/vol2/csit2421.pdf. 28 Maret 2015 (13.23 WIB).

Odeh, Ashraf, Aymen Abu-Errub, Qusai Shambour dan Nidal Turab. 2014. Arabic Text Categarization Algortithm Using Vector Evaluation Method. International Journal of Computer Science & Information Technology (IJCSIT) Vol. 6, No. 6. Jordan.

Saad, Motaz K. dan Wesam Ashor. 2010. Arabic Text Classification Using Decision Trees. http://site.iugaza.edu.ps/msaad/files/2011/01/mksaad-arabic-text-classification-using-decision-trees-CSIT2010.pdf. 28 Maret 2015 (13.11 WIB).

Schneider, Karl-Michael. 2005. Techniques For Improvind the Performance of Naïve Bayes for Text Classification. In Proceedings of CICLing, pages 682-693

Sun, Changqiu, Xiaolong Wang, dan Jun Xu. 2009. Study on Feature Selection in Finance Text Categorization. Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics. San Antonio.

Ting, S.L., W.H. Ip, dan Albert H.C. Tsang. 2011. Is Naïve Bayes a Good Classifier for Document Classification ?. International Journal of Software Engineering and It's Application. Vol.5, No.3, July. The Hong Kong Polytechnic University. Hongkong.

Trstenjak, Bruno, Sasa Mikac, dan Dzenana Donko. 2014. KNN with TF-IDF Based Framework for Text Categorization. Procedia Engineering 69 (2014) 1356 – 1364. Elsevier.

Wibisono, Yudi. 2005. Klasifikasi Berita Berbahasa Indonesia menggunakan Naïve Bayes Classifier. Seminar Nasional Matematika. UPI. Bandung.