Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin

Errissya Rasywir
Institut Teknologi Bandung
errissya.rasywir@gmail.com

Ayu Purwarianti
Institut Teknologi Bandung
ayu@stei.itb.ac.id

ABSTRAK
Klasifikasi berita hoax atau berita dengan informasi yang tidak benar merupakan salah satu aplikasi kategorisasi teks. Seperti aplikasi kategorisasi teks berbasis pembelajaran mesin pada umumnya, sistem ini terdiri atas praproses, ekstraksi fitur, seleksi fitur dan pengeksekusian model klasifikasi. Pada penelitian ini, eksperimen dilakukan untuk memilih teknik terbaik pada setiap sub proses dengan menggunakan 220 artikel berbahasa Indonesia dalam 22 topik (89 artikel hoax dan 131 artikel bukan hoax). Untuk praproses, hasil eksperimen terbaik dicapai oleh praproses tanpa *stemming* dan dengan penghapusan *stop word*. Untuk ekstraksi fitur, fitur unigram memiliki akurasi terbaik dibandingkan dengan bigram dan unigram+bigram. Untuk seleksi fitur, teknik terbaik adalah penggunaan operasi *union* pada *mutual information* dan *information gain.* Sedangkan untuk algoritma klasifikasi, dengan berbagai kombinasi di atas, algoritma *naïve bayes* menunjukkan hasil akurasi yang terbaik dibandingkan dengan SVM dan C4.5 dengan nilai akurasi 91.36%.

Kata Kunci
Artikel hoax, ekstraksi fitur, klasifikasi dokumen, seleksi fitur

1. PENDAHULUAN
Hoax adalah informasi sesat dan berbahaya karena menyesatkan persepsi manusia dengan menyampaikan informasi palsu sebagai kebenaran. Hoax mampu mempengaruhi banyak orang dengan menodai suatu citra dan kredibilitas (Chen et al, 2014). Hoax dapat bertujuan untuk mempengaruhi pembaca dengan informasi palsu sehingga pembaca mengambil tindakan sesuai dengan isi hoax. Sebagai pesan informasi palsu dan menyesatkan, hoax juga dapat menakut-nakuti orang yang menerimanya. Dengan demikian, sebaiknya hoax itu dapat dijelaskan, diidentifikasi dan diklasifikasikan (Petkovic et al, 2005). Penelitian terkait hoax sebelumnya telah dilakukan untuk domain email hoax (Petkovic et al, 2005; Vukovic et al, 2009; Chen et al, 2014). Hingga sekarang, belum ada studi dalam klasifikasi berita hoax.

Secara umum, sistem klasifikasi teks yang menggunakan pendekatan berbasis pembelajaran mesin terdiri dari praproses, ekstraksi fitur, seleksi fitur dan klasifikasi. Terdapat berbagai algoritma pembelajaran mesin pada klasifikasi teks yang dapat digunakan, contohnya *multinomial naive bayes* (Callum & Nigam,1998), model *multivariat bernoulli (*Joachim, 1997*), rocchio (*Joachim, 1997*), k-nearest neighbor* (Kotsiantis, 2005), dan *support vector machine* (Manning et al, 2008). Algoritma *Support Vector Machine* (SVM) adalah algoritma pembelajaran mesin terbaik untuk klasifikasi teks (Pilászy, 2005). Algoritma C4.5 adalah algoritma pembelajaran adaptif yang handal dan mampu menangani data *noise* (Maharani, 2009).

Teknik lain yang penting dalam sistem klasifikasi teks adalah seleksi fitur yang bertujuan untuk menghilangkan fitur *noise* yang mungkin menyebabkan klasifikasi menjadi tidak benar. Beberapa teknik seleksi fitur diantaranya adalah *document frequency thresholding* (Yang et al, 2003), *information gain (*Mitchell, 1996*), mutual information (*R. Fano, 1961*), term strength* (Yang et al, 2003) dan *chi-square* (T.E. Dunning, 1993). *Information gain* dan *chi-square* adalah metode yang paling efektif pada seleksi fitur untuk meningkatkan akurasi dari klasifikasi. *Term frequency* dapat meningkatkan nilai *recall* dalam *information retrieval*. Sementara itu, TFxIDF dapat berkontribusi untuk meningkatkan *recall* dan presisi (Langgeni et al, 2010). *Mutual information* adalah metode paling umum seleksi fitur yang digunakan dalam pemodelan bahasa statistik (Yang & Pedersen, 2003).

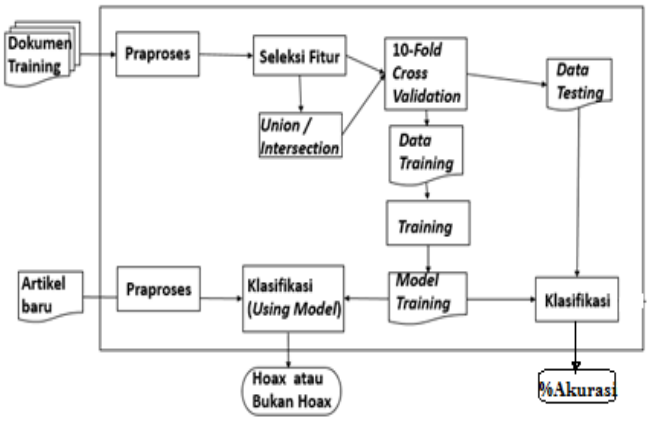
Untuk mendapatkan hasil yang terbaik, maka dibandingkan beberapa metode dalam membangun sistem klasifikasi berita hoax, baik dalam praproses, ekstraksi fitur, seleksi fitur dan algoritma pembelajaran mesin. Di sini, juga diusulkan operasi *union* dan *intersection* pada seleksi fitur untuk meningkatkan akurasi. Makalah ini terdiri dari analisis berita hoax berbahasa Indonesia, sistem klasifikasi, dan eksperimen.

2. ANALISIS BERITA HOAX
Fokus dalam penelitian ini adalah bagaimana menemukan fitur penciri pada proses klasifikasi dokumen berita hoax dalam bahasa Indonesia. Seperti dijelaskan di bagian sebelumnya yang mengatakan bahwa fitur penciri hoax pada email dan pesan teks dapat ditemukan dan diekstraksi dengan pola penulisan. Dalam email dan pesan teks (SMS) hoax, pola kata hoax dapat dikenali sebagai peringatan virus palsu, pesan berantai, permintaan bantuan palsu, pesan mengancam atau menakut-nakuti, petisi palsu, dan pernyataan bahwa pesan itu bukan hoax. Biasanya isi email dan pesan hoax bersifat *overstatements*, berlebihan dan bertujuan untuk menjual suatu produk.

Berbeda dengan email dan pesan hoax, dalam berita yang mengandung hoax, tidak ada pola yang dapat diidentifikasi. Dalam dokumen berita hoax, gaya penulisan bersifat bebas dan tidak kaku. Sulit bagi pembaca untuk membedakan mana berita hoax dan yang tidak. Salah satu cara untuk memeriksa apakah sebuah artikel merupakan berita hoax atau tidak adalah dengan melakukan klarifikasi terhadap berita yang sebenarnya. Jika klarifikasi dari artikel berita tidak tersedia, dapat dikatakan bahwa artikel tidak mengandung hoax karena tidak ada yang keberatan dengan isi berita. Penelitian ini mencoba untuk mencari informasi pada dokumen hoax untuk digunakan sebagai klasifikasi hoax otomatis. Sebagai data latih dan referensi, dilakukan pelabelan dokumen (hoax atau tidak hoax) secara manual.

3. KLASIFIKASI BERITA HOAX
Secara umum, langkah penelitian pada sistem klasifikasi berita hoax digambarkan pada Gambar 1. Penelitian ini dibagi menjadi dua tahap yaitu tahap pelatihan dan pengujian. Tahap pelatihan bertujuan membangun model klasifikasi terbaik, sementara tahap pengujian bertujuan untuk mengklasifikasikan dokumen input sebagai hoax atau tidak.

Pada pembangunan model, dilakukan pengujian 10-*fold cross validation* untuk memperoleh model terbaik. Model terbaik yang dihasilkan digunakan untuk mengklasifikasikan artikel baru sebagai dokumen masukan. Pada pembangunan model, terdapat 4 tahap utama yaitu praproses, ekstraksi fitur, seleksi fitur dan pelatihan. Setiap tahap dijelaskan pada uraian berikut.



Gambar 1. *Flowchart* Penelitian

3.1 Praproses Teks Berita Hoax

Semua dokumen baik data latih, data uji maupun artikel baru dikenai praproses terlebih dahulu. Tahap praproses terdiri dari pemrosesan leksikal dan pengubahan kata ke fitur kata. Pemrosesan leksikal bertujuan untuk memproses token kata dan terdiri atas modul tokenisasi, *case folding*, penghapuran *stopword* dan *stemming*. Modul *case folding* digunakan dengan asumsi bahwa informasi huruf kapital pada dokumen berita tidak mempengaruhi hasil klasifikasi teks. Modul penghapusan *stopword* digunakan untuk menghilangkan fitur kata yang tidak penting yang bisa mengganggu klasifikasi karena jumlahnya yang banyak pada dokumen. Modul *stemming* digunakan untuk menyamakan kata dengan lemma yang sama.

3.1.1 *Case Folding*

Case folding dilakukan untuk menghilangkan karakter selain huruf pada saat pengambilan informasi (Yates & Note, 1999). Proses ini melakukan pengubahan huruf dalam dokumen menjadi huruf kecil (‘a’ sampai ‘z’). Karakter selain huruf dianggap sebagai delimeter sehingga karakter tersebut dihapus dari dokumen. Tujuannya untuk menghilangkan *noise* pada saat pengambilan informasi (Yates & Note, 1999).

3.1.2 Tokenisasi

Proses tokenisasi adalah proses pemecahan kalimat menjadi kata atau frase (Yates & Note, 1999). Proses tokenisasi dalam penelitian ini memisahkan kata pada kalimat dengan menggunakan penanda spasi, lalu kata yang telah dipisah-pisah disimpan menjadi larik. Pada penelitian ini, proses tokenisasi yang dilakukan menghasilkan beberapa model fitur antara lain unigram, bigram dan gabungan keduanya.

3.1.3 Penghapusan *Stopword*

Penghapusan *stopwords* adalah proses pembuangan *stopwords* (kata yang sering muncul dan tidak dipakai). Proses ini bertujuan untuk mengurangi volume kata. *Stopwords* dapat berupa kata depan, kata penghubung, dan kata pengganti. Contoh *stopwords* dalam bahasa Indonesia adalah “yang”, “ini”, “dari”, “ ke“, “di”, “dari” (Yates & Note, 1999). Pengaruh tidaknya penghilangan *stopwords* tergantung pada jenis klasifikasi dan data yang terkumpul. Berdasarkan hal tersebut, maka dilakukan pengujian penghapusan *stopword* pada dokumen teks berita hoax.

3.1.4 *Stemming*

Stemming adalah proses pemotongan imbuhan atau pengembalian kata berimbuhan menjadi kata dasar (Yates & Note, 1999). Proses ini bertujuan untuk mengurangi variasi kata yang sebenarnya memiliki kata dasar yang sama. Algoritma pemotongan imbuhan dalam penelitian ini menggunakan algoritma Adriani dan Nazief. Pada penelitian ini, juga dilakukan pengujian pada praproses *stemming* ini untuk melihat pengaruh penggunaan kata dasar pada klasifikasi berita hoax. Sehingga dengan demikian data pengujian *stemming* yang diperoleh adalah 4 jenis yakni:

a. Dokumen *non-stemming* + tanpa penghapusan *stopword*

b. Dokumen *stemming* + tanpa penghapusan *stopword*

c. Dokumen *non-stemming* + penghapusan *stopword*

d. Dokumen s*temming* + penghapusan *stopword*

3.2 Ekstraksi Fitur Berita Hoax

Ekstraksi fitur adalah proses mengekstrak seluruh fitur kata yang terdapat dalam dokumen latih. Keluaran dari proses ini adalah kumpulan kata yang dijadikan penciri dokumen berita hoax dan bukan hoax. Fitur kata ini diperoleh dari hasil tokenisasi. Dimana setiap kata terpisah berdasarkan penanda spasi, lalu kata disimpan menjadi larik. Pada penelitian ini, proses ekstraksi fitur kata yang dilakukan menghasilkan beberapa model fitur antara lain:

1. Unigram
2. Bigram
3. Gabungan Unigram & Bigram

Ketiga jenis fitur tersebut diujikan pada seluruh model pengujian praproses, pengujian seleksi fitur dan algoritma klasifikasi (*classifier)*.

3.3 Seleksi Fitur Berita Hoax

Hasil fitur yang diperoleh selanjutnya diseleksi untuk mengambil hanya sejumlah fitur yang diasumsikan memegang informasi penting dari kelas dokumen. Teknik yang digunakan untuk seleksi fitur adalah *information gain* (IG), *mutual information* (MI), *chi-square* (CS), *term frequency* (TF) dan TFxIDF. Dalam eksperimen, dilakukan perbandingan antara tanpa seleksi fitur, dengan seleksi fitur tunggal, dan dengan gabungan seleksi fitur.

3.3.1 *Information Gain*

Teknik seleksi fitur dengan *Information gain* artinya adalah memilih simpul fitur dari pohon keputusan berdasar nilai *information gain*. Nilai *information gain* sebuah fitur diukur dari pengaruh fitur tersebut terhadap keseragaman kelas pada data yang dipecah menjadi sub data dengan nilai fitur tertentu. Keseragaman kelas (entropy) dihitung pada data sebelum dipecah dengan persamaan (1) dan pada data setelah dipecah dengan persamaan (2) (Mitchell, 1996):

$$Entropy\ (S) = \sum_{i=1}^{k}(P_i) \log 2\ (P_i) \quad (1)$$

Dengan nilai P_i adalah proporsi data S dengan kelas i. K adalah jumlah kelas pada output S.

$$Entropy\ (S, A) = \sum_{i=1}^{v}(\frac{Sv}{S} * Entropy\ (Sv)) \quad (2)$$

Dengan nilai v adalah semua nilai yang mungkin dari atribut A, S_v adalah subset dari S dimana atribut A bernilai v. Nilai *information gain* dihitung dari persamaan berikut:

$$Gain\ (S, A) = Entropy\ (S) - Entropy\ (S, A) \quad (3)$$

Dengan nilai *Gain* (S,A) adalah nilai *information gain*. *Entropy* (S) adalah nilai *entropy* sebelum pemisahan. *Entropy* (S, A) adalah nilai *entropy* setelah pemisahan. Besarnya nilai *information gain* menunjukkan seberapa besar pengaruh suatu atribut terhadap pengklasifikasian data (Mitchell, 1996).

3.3.2 *Mutual Information*

Mutual Information (MI) menunjukkan seberapa banyak informasi mengenai ada atau tidaknya sebuah kata memberikan konstribusi dalam membuat keputusan klasifikasi secara benar atau salah. Nilai dari MI disimbolkan dengan notasi I (R. Fano, 1961):

$$I\ (U; C) = \sum_{ec\in\{1,0\}} P(U = et, C = ec) log_2 \frac{P\ (U=et, C=ec)}{P(U=et)P\ (C=ec)} \quad (4)$$

Dengan, U adalah variabel acak dengan nilai-nilai et. Nilai et =1 adalah dokumen berisi kata t. Nilai et = 0 adalah dokumen yang tidak mengandung kata t. Nilai C adalah variabel acak dengan nilai-nilai ec. Nilai ec = 1 adalah dokumen dikelas c. Nilai ec = 0 adalah dokumen tidak dikelas c.

3.3.3 *Chi-Square*

Chi-square menguji hipotesis perbandingan antara frekuensi data observasi dengan frekuensi harapan (ekspektasi) berdasarkan pada suatu hipotesis tertentu untuk setiap kasus atau data. Perhitungan *chi-square* adalah sebagai berikut (T.E. Dunning, 1993):

$$x^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \quad (5)$$

Dengan nilai o_i adalah frekuensi observasi dan e_i adalah frekuensi ekspektasi. Lalu hitung:

$$e_i = \frac{\sum fk\ x\ \sum fb}{\sum T} \quad (6)$$

Dengan nilai Σfk adalah jumlah frekuensi pada kolom. Nilai Σfb adalah jumlah frekuensi pada baris. Nilai ΣT adalah jumlah keseluruhan baris atau kolom. *Chi-square* menguji hubungan atau pengaruh dua variabel dan mengukur keterkaitan antara variabel satu dengan lainnya. Penghitungan nilai *chi-square* pada setiap kata t yang muncul pada setiap kelas c dapat dibantu dengan menggunakan tabel kontingensi (Tabel 1). Nilai pada tabel kontingensi merupakan nilai frekuensi observasi dari suatu kata terhadap kelas (T.E. Dunning, 1993).

Table 1. Kontigensi Kata

Kata	Kelas	
	1	0
1	A	B
0	C	D

Penghitungan tersebut dapat disederhanakan menjadi:

$$X^2\ (\mathrm{t}, \mathrm{c}) = \frac{N\ (AD-CB\ )^2}{(A+C)(B+D)(A+B)(C+D)} \quad (7)$$

Dengan nilai t adalah kata yang diujikan terhadap kelas c. Nilai N adalah jumlah dokumen latih. Nilai A adalah banyaknya dokumen kelas c yang memuat kata t. Nilai B adalah banyaknya dokumen yang tidak berada di kelas c namun memuat kata t. Nilai C adalah banyaknya dokumen yang berada di kelas c namun tidak memiliki kata t di dalamnya. Nilai D adalah banyaknya dokumen yang bukan merupakan dokumen kelas c dan tidak memuat kata t (T.E. Dunning, 1993).

3.3.4 *Term Frequency* (TF)

Term frequency atau frekuensi kata dapat dinyatakan dalam notasi rumus berikut (Yang & Liu, 1999):

$$\mathrm{tf}(\mathrm{t}, \mathrm{d}) = \sum_{x\in d} fr\ (x, t) \quad (8)$$

Dengan nilai tf adalah jumlah kemunculan setiap kata t, nilai fr(x,t) adalah fungsi sederhana yang didefenisikan sebagai $fr\ (x, t)$, dengan:

$$r\ (x, t) = \begin{cases} 1, if\ x = t \\ 0,\ Lainnya \end{cases} \quad (9)$$

Penggunaan TF dapat memperbaiki nilai *recall* pada *information retrieval.* Hal ini disebabkan TF muncul di banyak teks, sehingga kata tersebut memiliki keunikan yang kecil. Untuk memperbaiki permasalahan ini, kata dengan nilai frekuensi yang tinggi sebaiknya dibuang dari set kata. Penggunaan *threshold* yang optimal dapat membantu seleksi fitur (Langgeni et al, 2010).

3.3.5 *Term Frequency - Inversed Document Frequency* (TFxIDF*)*

Berikut adalah rumus menghitung TFxIDF (Yang & Liu, 1999):

$$tf_{t,d}\ x\ idf \quad (10)$$

Dengan nilai tf adalah jumlah kemunculan kata t. Nilai d adalah dokumen. Nilai $tf_{t,d}\ x\ idf$ adalah kemunculan kata t setiap dokumen. Nilai idf merupakan kemunculan kata t pada semua dokumen (pembobotan global):

$$idf = \log \frac{N}{df_t} \quad (11)$$

Dengan nilai N adalah banyaknya dokumen, Nilai df_t adalah jumlah dokumen yang mengandung kata t.

3.4 Pelatihan dengan Menggunakan Algoritma Pembelajaran Mesin

Pelatihan atau pembangunan model dilakukan setelah tahap seleksi fitur. Adapun algoritma pembelajaran mesin yang dipilih pada sistem klasifikasi berita hoax ini adalah *naïve bayes*, SVM (*Support Vector Machine*) dan Algoritma C4.5. Algoritma *naïve bayes* dipilih dalam penelitian ini karena telah terbukti efektif untuk kategorisasi teks, prosesnya sederhana, cepat dan akurasi klasifikasi yang tinggi (Dai, 1997). Algoritma *Support Vector Machine* (SVM) dipilih karena merupakan algoritma pembelajaran mesin terbaik untuk klasifikasi teks (Pilászy, 2005). Pemilihan algoritma

C4.5 karena merupakan algoritma pembelajaran adaptif yang handal dan mampu menangani data *noise* (Han & Kamber, 2006).

3.4.1 *Naïve Bayes*

Untuk merepresentasikan sebuah kelas dokumen, terdapat karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi yang berguna untuk menjelaskan bahwa peluang masuknya sampel karakteristik tertentu kedalam kelas posterior. Klasifikasi *naive bayes* diasumsikan bahwa ada atau tidaknya ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. Persamaan dari teorema *bayes* adalah (Mitchell, 1997):

$$P(H|X)=(P(X|H)P(H))/(P(X)) \quad (12)$$

Di mana nilai X adalah data kelas yang belum diketahui, H adalah hipotesis X pada label tertentu, P(H|X) adalah probabilitas H berdasarkan kondisi X (*posteriori*), P(H) adalah probabilitas H (*prior*), P(X|H) adalah probabilitias X pada hipotesis H, P(X) adalah probabilitas X.

3.4.2 *Support Vector Machine* (SVM)

SVM merupakan algoritma pembelajaran mesin universal (dapat menangani berbagai jenis data) yang memanfaatkan fungsi pembatas linear sebagai basisnya (Joachims, 1998). Namun, tidak semua data dapat dipisahkan secara linear dalam dua dimensi. Oleh karena itu, fungsi pembatas linear tersebut kemudian ditransformasi menjadi *hyperplanes* dengan menggunakan fungsi kernel sehingga *hyperplanes* tersebut dapat memisahkan data dalam ruang dimensi yang lebih tinggi (Hoffman, 2006).

Pada SVM, fungsi pemisah bertujuan untuk menentukan kelas. Bidang pemisah pendukung dari kelas +1 dan bidang pemisah pendukung dari kelas −1. Secara matematika, mencari bidang pemisah terbaik ekuivalen dengan memaksimalkan margin antara dua kelas. Memaksimalkan margin antara kedua kelas sama dengan meminimumkan fungsi tujuan dengan memperhatikan pembatas. Bidang pemisah terbaik ialah bidang pemisah yang menghasilkan nilai margin terbesar dan berada di tengah-tengah antara dua set objek dari dua. Nilai margin merupakan jarak antara bidang pemisah dengan elemen terluar dari kedua kelas. Dalam hal ini fungsi pemisah yang dicari adalah fungsi linear sebagai berikut (Manning dkk, 2008):

$$f(x) = sign\ (w^T\ x_i + b = 0) \quad (13)$$

Dengan nilai W adalah bobot yang merepresentasikan posisi *hyperplane* pada bidang normal, X adalah vektor data masukan. B adalah bias yang merepresentasikan posisi bidang relatif terhadap pusat koordinat.

Selanjutnya data dikelompokkan dengan menggunakan fungsi pemisah yang sudah ditemukan, di mana untuk menentukan kelasnya: Fungsi w.xi + b = +1 adalah bidang pemisah pendukung dari kelas +1. Fungsi w.xi + b = −1 adalah bidang pemisah pendukung dari kelas −1. Bidang pemisah terbaik ekuivalen dengan memaksimalkan margin antara dua kelas yang dihitung dengan formula 2‖w‖ 2. Memaksimalkan margin antara kedua kelas sama dengan meminimumkan fungsi tujuan 1 2‖w‖ 2 dengan memperhatikan pembatas yi(w·xi+b) ≥1 dengan xi adalah data input dan yi adalah keluaran dari data xi. Selanjutnya, masalah klasifikasi diformulasikan ke dalam *quadratic programming* (QP) yang diselesaikan dengan *lagrange multiplier* (Manning dkk, 2008):

$$L\ (w, b, \alpha) = \frac{1}{2}\ ||w||^2\text{-}\sum_{i=1}^{N} \alpha_i y_i\ (w.x_i + b - 1) \quad (14)$$

Dengan nilai ∝I adalah *lagrange multiplier* yang berkorespondensi dengan xi.

3.4.3 Algoritma C4.5

Algoritma ini dapat menangani data kontinu dan diskrit. Untuk pemisahan obyek atribut kontinu, pengurutan berdasarkan atribut, kemudian membentuk minimum ambang dari contoh yang ada dari kelas mayoritas pada setiap partisi yang bersebelahan, lalu digabungkan partisi yang bersebelahan tersebut dengan kelas mayoritas yang sama. Jika suatu dataset mempunyai *missing value*, maka atribut tersebut diganti dengan nilai rata-rata variabel yang bersangkutan. Dalam pemisahan obyek dilakukan tes terhadap atribut dengan mengukur tingkat ketidakmurnian pada sebuah simpul (Larose, 2005).

Pohon keputusan dibangun dengan menghitung nilai Information Gain dari setiap fitur sebagai isi dari setiap simpul. Persamaan yang digunakan untuk menghitung *information gain* tsb dapat dilihat pada persamaan (1), (2) dan (3). Pembentukan simpul dilakukan terus hingga diperoleh data yang seragam untuk sebuah simpul. Selanjutnya dilakukan pemotongan pohon (*pruning*) untuk menghindari *overfitting*.

4. EKSPERIMEN

4.1 Data dan Skenario Eksperimen

Tujuan dari skenario eksperimen yang telah dilaksanakan pada penelitian ini mencakup:

1. Evaluasi penggunaan modul *stopwords elimination* dan *stemming* pada praproses
2. Pemilihan jenis fitur (unigram, bigram dan gabungan unigram-bigram)
3. Pemilihan teknik seleksi fitur (tanpa seleksi fitur, 5 jenis seleksi fitur, *union* dan *intersection* dari seleksi fitur)
4. Pemilihan algoritma pembelajaran mesin terbaik

Dataset yang digunakan untuk ektraksi dan diseleksi fiturnya berjumlah 220 artikel (89 artikel hoax dan 131 bukan hoax). Kumpulan dataset digunakan untuk semua skenario eksperimen. Untuk dataset tanpa operasi *union* dan *intersection* berisi skenario eksperimen dengan parameter 4 kombinasi praproses (*stopword elimination*, *non stopword elimination*, *stemming*, *non stemming*) x 3 jenis fitur (unigram, bigram, gabungan) x 5 jenis seleksi fitur (*information gain* (IG), *mutual information* (MI), *chi-square* (CS), *term frequency* (TF) dan TFxIDF). Sehingga untuk eksperimen dengan seleksi fitur tunggal, terdapat 60 jenis dataset.

Sedangkan dataset dengan operasi *union* dan *intersection* pada seleksi fitur ditunjukkan oleh Tabel 2 berikut.

Tabel 2. Fitur dengan Operasi *Union* dan *Intersection*

Intersection	***Union***
IG ∩ MI	IG ∪ MI
IG ∩ CS	IG ∪ CS
IG ∩ TF	IG ∪ TF
IG ∩ TFxIDF	IG ∪ TFxIDF
MI ∩ CS	MI ∪ CS

Intersection	*Union*
MI ∩ TF	MI ∪TF
MI ∩ TFxIDF	MI ∪ TFxIDF
CS ∩ TF	CS ∪ TF
CS ∩ TFxIDF	CS ∪ TFxIDF
TF ∩ TFxIDF	TF ∪ TFxIDF

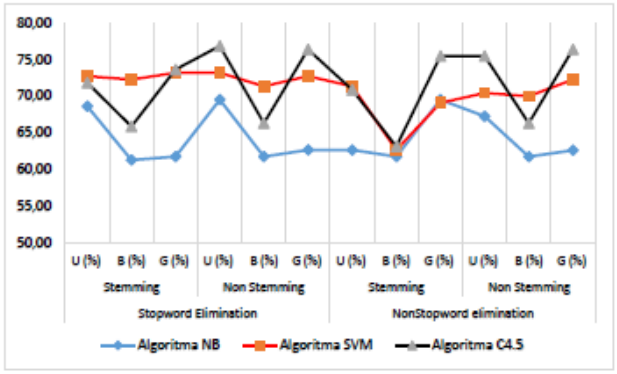
Eksperimen pada penelitian ini dilakukan pada sistem operasi *windows* 7 dengan 32 bits *system*, IDE Netbeans dan MySQL untuk *database*. Sistem dibangun dengan bahasa pemrograman JAVA.

4.2 Hasil Eksperimen

Pengujian penelitian menggunakan model 10-*folds cross validation*. Untuk perhitungan akurasi hasil pengujian dilakukan secara statistik yakni total *instance* yang benar dibagi total seluruh *instance* (setiap dataset terdiri dari 220 *instance*) dikali 100%. Seluruh pengujian klasifikasi yang dilakukan meliputi pengujian praproses (*stemming* & *stopword elimination*) dengan 3 jenis *classifier* yang telah disebutkan. Namun, dari hasil kombinasi pengujian tersebut terbagi 2 jenis pengujian utama yakni klasifikasi tanpa dan dengan seleksi fitur. Untuk pengujian klasifikasi dengan seleksi fitur dikombinasikan dengan operasi *union* dan *intersection*.

4.2.1 Hasil Klasifikasi Tanpa Seleksi Fitur

Gambar 2 menunjukkan eksperimen dengan tanpa seleksi fitur. Dari Gambar 2, dapat dilihat bahwa modul praproses terbaik adalah dengan menggunakan penghapusan *stopword* dan tanpa *stemming*. Sedangkan jenis fitur terbaik adalah unigram dan gabungan unigram-bigram.



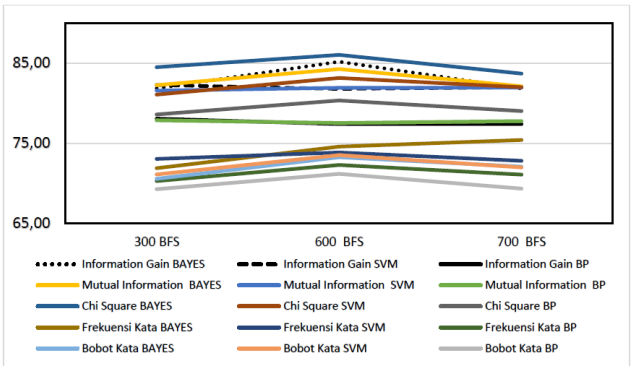
Gambar 2. Hasil Klasifikasi Tanpa Seleksi Fitur

Hasil klasifikasi tanpa seleksi fitur ini memberikan nilai yang tidak lebih dari 80 %.

4.2.2 Hasil Klasifikasi Dengan Seleksi Fitur

4.2.2.1 Seleksi Fitur Tunggal

Gambar 3 berikut menunjukkan eksperimen dengan seleksi fitur yaitu *information gain*, *mutual information*, *chi-square*, *term frequency* dan TFxIDF. Hasil klasifikasi tertinggi diperoleh dengan jumlah fitur 600 sebagai *Best First Selection* / BFS (dari 3 jenis pengujian pada jumlah fitur yakni 300, 600 & 700 BFS).

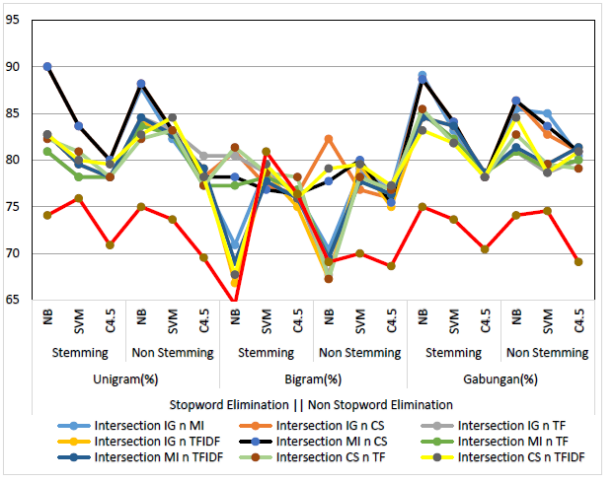


Gambar 3. Hasil Klasifikasi Dengan Seleksi Fitur Tunggal

Berdasarkan hasil Gambar 3,maka eksperimen klasifikasi dengan seleksi fitur selanjutnya diujikan dengan jumlah fitur sebesar 600 fitur kata.

4.2.2.2 Hasil Klasifikasi Dengan Seleksi Fitur Menggunakan *Intersection & Union*

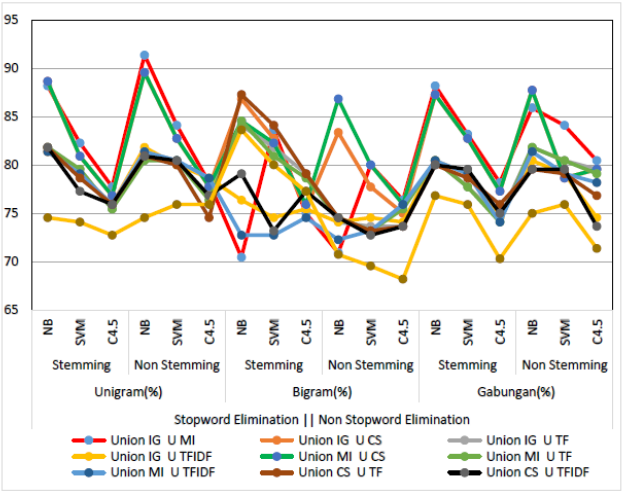
Gambar 4 menunjukkan eksperimen pada seleksi fitur yang dikombinasikan dengan operasi irisan atau *intersection*. Seleksi fitur ini berasal dari 5 jenis seleksi fitur yang telah ditetapkan yakni *information gain*, *mutual information*, *chi-square*, *term frequency* dan TFxIDF. Untuk daftar kombinasi seleksi fiturnya dapat dilihat pada tabel 2 (kolom *Intersection*). Pada gambar 5 hasil tertinggi diperoleh dari fitur hasil irisan antara *mutual information* dan *chi-square* (MI ∩ CS) dengan akurasi 90 %. Nilai tersebut adalah hasil *classifier naïve bayes* dengan model fitur unigram yang telah di *stemming* dan mengalami *stopword elimination*.



Gambar 4. Hasil Klasifikasi Dengan Seleksi Fitur Dengan Operasi *Intersection*

4.2.3 Perbandingan Algoritma Klasifikasi

Gambar 5 berikut menunjukkan eksperimen pada seleksi fitur yang dikombinasikan dengan operasi gabungan atau *union*. Untuk daftar kombinasi seleksi fiturnya dapat dilihat pada tabel 2 (kolom *Union*). Pada gambar 5, hasil tertinggi diperoleh dari fitur hasil gabungan (*union*) antara *mutual information* dan *information gain* (IG ∪ MI) dengan akurasi 91.36 %. Nilai tersebut adalah hasil *naïve bayes* dengan model fitur unigram yang tidak di-*stemming* dan namun sudah mengalami penghilang *stopword*.



Gambar 5. Hasil Klasifikasi Seleksi Fitur Dengan Operasi *Union*

Gambar 6 adalah hasil klasifikasi dari pengujian tertinggi yakni dengan penggunaan seleksi fitur yang dioperasikan secara *union* atau *intersection*. Hasil pengujian ini di kelompokkan berdasarkan penggunaan *classifier*. Dari gambar 6 tersebut dapat dilihat bahwa algoritma *naïve bayes* mampu melakukan klasifikasi berita hoax paling baik dibanding algoritma lainnya (SVM dan C4.5).

Gambar 6. Perbandingan Algoritma Klasifikasi

4.2.4 Analisis Hasil Keseluruhan Pengujian

Dari seluruh hasil klasifikasi yang diperoleh sebelumnya dapat diperoleh beberapa kesimpulan yang dapat dilihat pada tabel 3, 4 dan 5.

Tabel 3. Seleksi Fitur Yang Menghasilkan Akurasi tertinggi

Fitur	(%)	*Union*	(%)	*Intersection*	(%)
IG	90.45	IG ∪ MI	91.36	IG ∩ MI	90
MI	89.55	IG ∪ CS	89.55	IG ∩ CS	90
CS	88.63	MI ∪ CS	89.55	MI ∩ CS	90

Tabel 3 adalah nilai klasifikasi tertinggi yang dihasilkan seluruh pengujian seleksi fitur. Kolom 'fitur' adalah seleksi fitur yang digunakan secara tunggal. Seleksi fitur tersebut adalah *information gain* (IG), *mutual information* (MI) dan *chi-square* (CS). Tabel tersebut menunjukan bahwa klasifikasi dengan akurasi tertinggi adalah hasil *union* (IG dan MI) yakni sebesar 91.36%. Pada tabel tersebut terdapat beberapa penggunaan seleksi fitur berbeda namun nilai akurasinya sama.

Tabel 4. Kombinasi 3 Seleksi Fitur Terbaik

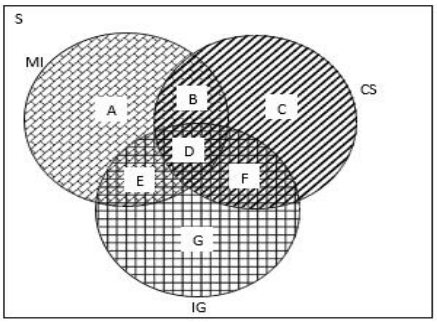
Kombinasi lainnya	(%)
IG ∪ MI ∪ CS	89.55
IG ∩ MI ∩ CS	90.45
(IG ∩ MI) ∪ (IG∩ CS) ∪ (MI ∩ CS)	91.36

Tabel 4 diperoleh berdasarkan nilai tertinggi, disini dilakukan kombinasi dari 3 jenis seleksi fitur terbaik yakni IG, MI dan CS. Kombinasi ini untuk melihat bagaimana jika ketiga hasil fitur tersebut digabungkan atau diiriskan. Nilai akurasi diperoleh pada tabel 4 mempunyai kesamaan nilai akurasi pada tabel 3. Namun, penggabungan atau irisan ketiganya tidak menghasilkan nilai yang lebih baik.

Tabel 5. Akurasi Yang Sama Dengan Notasi Berbeda

Seleksi fitur Akurasi Sama		
Notasi I	**Notasi II**	**(%)**
IG ∪ MI ∪ CS	IG∪CS=MI∪CS	89,55
IG ∩ MI ∩ CS	IG	90,45
(IG∩MI)∪(IG∩CS) ∪(MI ∩CS)	IG ∪ MI	91,36
IG ∩ MI	IG∩CS= MI∩CS	90

Tabel 5 menunjukkan kombinasi seleksi fitur berbeda dengan akurasi sama. Hal ini dilakukan untuk analisis terhadap penggunaan seleksi fitur pada klasifikasi berita hoax. Hal ini menunjukkan 3 seleksi fitur terbaik yang menghasilkan nilai akurasi tertinggi adalah *information gain, mutual information* dan *chi-square*. Diagram venn pada gambar 7 merupakan representasi fitur yang berada pada hasil seleksi fitur (IG, MI dan CS). Diagram tersebut menunjukkan hubungan dari ketiga seleksi fitur.



Gambar 7. Hubungan Seleksi Fitur IG, MI dan CS

Gambar 7 menghasilkan himpunan:

- (IG ∩ MI) ∪ (IG ∩ CS) ∪ (MI ∩ CS) = {B, D, E, F} -> 91.36 %
- IG ∪ MI = {A, B, D, E, F, G} -> 91.36 %
- IG ∪ MI ∪ CS = {A, B, C, D, E, F, G} -> 89.55%
- IG ∪ CS = { B, C, D, E, F, G} -> 89.55 %
- MI ∪ CS = { A, B, C, D, E, F} -> 89.55 %
- IG ∩ MI ∩ CS = {D} -> 90,45 %
- IG = {D, E, F, G } -> 90,45 %
- IG ∩ MI = {D, E} -> 90 %
- IG ∩ CS = {D, F} -> 90 %
- MI ∩ CS = { B, D } -> 90 %

Himpunan {IG ∪ MI ∪ CS}, {IG ∪ CS} dan {MI ∪ CS} menghasilkan 89.55 %, namun {IG ∩ MI ∩ CS} menghasilkan 90.45 %. Hal ini menunjukkan bahwa {IG ∪ MI ∪ CS} mempunyai fitur yang tidak efisien {A, C, G}. Notasi {IG ∪ MI} dan {(IG ∩ MI) ∪ (IG ∩ CS) ∪ (MI ∩ CS)} menghasilkan akurasi sama sebesar 91.36%, namun fiturnya berbeda. Dilihat dari hasil penguraian fitur pada himpunan tersebut terdapat fitur {A} dan {G}. Hal ini

menunjukan bahwa fitur tersebut relevansinya sama-sama tidak berpengaruh terhadap kelas hasil klasifikasi artikel hoax. Fitur {A} adalah fitur milik *mutual information* (MI) dan fitur {G} milik *information gain* (IG) menunjukan bahwa pengaruh fitur yang dihasilkan dan terhadap kelas seimbang. Fitur {B}, {E} dan {F} juga memberikan pengaruh yang setara dimana keberadaan ketiga mempunyai pengaruh yang sama terhadap hasil klasifikasi (dengan akurasi 90 %). Fitur {B} adalah *intersection* MI dan CS, fitur {E} adalah *intersection* antara MI dan IG, sedangkan fitur {F} adalah *intersection* antara IG dan CS. Untuk fitur paling kuat adalah fitur {D} yakni hasil *intersection* ketiganya. Visualisasi menunjukkan bahwa relevansi fitur yang dihasilkan IG dan MI setara, sedangkan dengan keberadaan fitur {C} yang dapat menurunkan akurasi, menunjukkan bahwa relevansi CS terhadap kelas dibawah IG dan MI.

4.3 Analisis Hasil Eksperimen

4.3.1 Analisis Perbandingan Seleksi Fitur

Hasil pengujian menyatakan bahwa hasil terbaik dihasilkan oleh seleksi fitur berbasiskan probabilitas. Teknik *information gain*, *mutual information* dan *chi-square* sebagai fitur seleksi berbasiskan probabilitas mampu menghasilkan nilai akurasi rata-rata 81.82% (rata-rata=total akurasi 108 pengujian menggunakan seleksi fitur IG, MI dan CS dibagi 108), sedangkan *term frequency* dan TFxIDF sebagai seleksi fitur berbasiskan frekuensi menghasilkan 73.42% (rata-rata = total akurasi 72 pengujian yang menggunakan seleksi fitur TF dan TFxIDF dibagi 72). Dengan demikian, dapat disimpulkan bahwa seleksi fitur berbasiskan probabilitas mencapai akurasi yang lebih baik untuk klasifikasi artikel hoax daripada fitur berdasarkan frekuensi.

Fitur kata yang dihasilkan oleh seleksi fitur berbasis probabilitas (IG, MI & CS) tersebut dapat dilihat pada tabel 6 dibawah ini. Pada tabel 6 ini ditampilkan hanya 13 fitur dengan bobot tertinggi hasil seleksi fitur tersebut.

Tabel 6. Tabel Fitur Kata Hasil Seleksi Fitur

Dataset: unigram Praproses: *non stemming- stopword,elimination, Classifier: naive bayes*			
IG U MI	**IG**	**MI**	**CS**
(13 dari 625 BFS; 91.36%)	**(13 dari 600 BFS; 90.45%)**	**(13 dari 600 BFS; 89.55%)**	**(13 dari 600 BFS; 88.63%)**
hoax	hoax	hoax	hoax
beredar	beredar	beredar	beredar
berita	berita	berita	berita
com	com	com	com
terbang	terbang	terbang	terbang
kebiasaan	kebiasaan	kebiasaan	jerman
sosial	sosial	sosial	sosial
jerman	jerman	jerman	kebiasaan
belaka	belaka	belaka	kabar
buruk	buruk	buruk	warga
memasang	memasang	memasang	anak
www	www	www	media
kabar	kabar	kabar	www

*BFS: Best First Selection; IG U MI: Information Gain yang di-union-kan dengan Mutual Information; IG: Information Gain; MI: Mutual Information; CS: Chi-Square

Kolom IG U MI pada tabel 6 di atas, berisi 13 fitur kata tertinggi hasil seleksi fitur *information gain* yang di-*union*-kan dengan *mutual information*. Fitur-fitur kata pada kolom tersebut mampu menghasilkan akurasi klasifikasi hingga 91.36%. Kolom IG pada tabel 6 di atas, berisi 13 fitur kata tertinggi hasil seleksi fitur *information gain* dengan akurasi mencapai 90.45%. Kolom MI pada tabel 6 di atas, berisi 13 fitur kata tertinggi hasil *mutual information* dengan akurasi mencapai 89.55%. Kolom CS pada tabel 6 di atas, berisi 13 fitur kata tertinggi hasil *chi-square* dengan akurasi mencapai 88.63%. Fitur kata yang digunakan adalah jenis unigram yang tidak mengalami *stemming* namun telah dieliminasi *stopwords*-nya. Akurasi tersebut diperoleh dari hasil klasifikasi oleh *naïve bayes*.

Dari daftar fitur kata pada tabel 6 dapat dilihat bahwa terdapat kesamaan dan perbedaan fitur kata di setiap kolom berbeda. Hal itu dipengaruhi oleh metode seleksi fitur yang digunakan. Operasi *union* antara IG dan MI mampu menyeleksi fitur paling baik dibanding penggunaan secara tunggal pada seleksi fitur berbasis probailitas.

4.3.2 Analisis *Incorrect* % Akurasi Tertinggi

Nilai *incorrect* % merupakan hasil pemeriksaan *instance* yang salah dari akurasi tertinggi pada pengujian menggunakan 10-*folds* CV. Dimana hasil akurasi tertinggi adalah hasil klasifikasi menggunakan seleksi fitur dengan operasi *union* antara *information gain* dan *mutual information* yakni sebesar 91.36%. Pada pengujian tersebut diperiksa *instance* yang salah diklasifikasikan. Pemeriksaan menunjukkan bahwa penyebab kegagalan klasifikasi berita hoax adalah bahwa terdapat *instance* yang hasil perhitungan $\sum P(C|W)$ dan *voting* setiap *instance* lebih besar kecenderungannya untuk masuk ke suatu kelas tertentu, namun kelas yang telah dimiliki *instance* sebelumnya tidak sama dengan hasil klasifikasi. Sehingga *instance* terklasifikasi salah.

Hal itu disebabkan hasil perhitungan probabilitas *instance* terhadap seluruh topik dan topik sejenis menunjukkan hasil berbeda. Sebanyak 72.22 % (13/18) *instance* hasil probabilitas terhadap seluruh topik menunjukan nilai sama dengan hasil klasifikasi. Namun, hanya 44.44 % (8/18) *instance* hasil probabilitas terhadap topik sejenis yang menunjukkan hasil yang sama dengan hasil klasifikasi. Hal ini menunjukkan bahwa nilai probabilitas terhadap seluruh topik lebih menentukan hasil klasifikasi dibandingkan dengan probabilitas topik sejenis.

4.3.3 Nilai F1 Dari Hasil Akurasi Tertinggi

Hasil tertinggi dari seluruh pengujian adalah hasil klasifikasi menggunakan seleksi fitur dengan kombinasi operasi *union* antara *information gain* dan *mutual information*. Dataset tersebut diklasifikasikan menggunakan *naive bayes* dengan pengujian 10-*fold cross validation*. Nilai yang diketahui antara lain *false postive* (FP), *false negative* (FN), *correct* %, *confusion matrix* dan F1.

Tabel 7. Hasil Akurasi Tertinggi

fold	***Correct* %**	***Incorrect* %**	**FP**	**FN**	**F1**
1	90.90	9.09	2	0	0.928
2	95.45	4.54	1	0	0.962
3	90.90	9.09	2	0	0.928
4	100	0	0	0	1
5	100	0	0	0	1
6	90.90	9.09	1	1	0.923
7	81.81	18.18	3	1	0.857
8	86.36	13.63	2	1	0.888
9	77.27	22.72	3	2	0.814
10	100	0	0	0	1

Pada tabel 6 dapat dilihat bahwa nilai F1 terbaik dihasilkan model *fold*-4, *fold*-5 dan *fold*-10 dengan nilai 100%. Dan nilai F1 terendah adalah 0.814 yang dihasilkan model *fold*-9. Nilai F1 yang rendah ini karena total FP dan FN yang dihasilkan klasifikasi. Dataset dengan nilai FP dan FN tertinggi dihasilkan *fold*-9. Tingginya nilai FP dan FN disebabkan ketidak-konsistenan keberadaan fitur dalam sebuah kelas serta kecenderungan fitur suatu kelas berbeda dengan keberadaan

kelas fitur itu berada. Oleh karena itu, model yang dihasilkan dapat dikatakan belum konsisten karena menghasilkan 14 FP dan 5 FN.

Pada modul praproses, penghilangan *stopword* memberikan hasil akurasi yang lebih tinggi secara rata-rata 1.13%, sesuai dengan asumsi bahwa penghilangan kata yang tidak penting dapat meningkatkan akurasi. Namun, modul *stemming* ternyata memberikan akurasi lebih rendah. Berikut beberapa poin kesimpulan yang dihasilkan seluruh pengujian:

1. Fitur tanpa *stemming* memberikan akurasi lebih baik. Ini menunjukkan bahwa penciri berita hoax ditentukan secara leksikal.
2. Model fitur terbaik adalah unigram. Namun, gabungan antara unigram dan bigram memberikan hasil yang lebih baik secara rata-rata.
3. Algoritma klasifikasi yang menghasilkan nilai akurasi terbaik adalah *naive bayes* dan terendah adalah C4.5 dengan pengujian 10-*fold cross validation.*
4. Seleksi fitur dengan operasi *union* antara *information gain* dan *mutual information* menghasilkan nilai tertinggi sebesar 91.36%. *Information gain* saja menghasilkan 90.45%. Operasi *intersection* tertinggi menghasilkan 90%. Hasil tersebut lebih baik dibandingkan klasifikasi tanpa seleksi fitur yakni rata-rata sebesar 69.57%.
5. Probabilitas kata terhadap seluruh topik (22 topik) menghasilkan nilai yang lebih baik dibandingkan topik sejenis.
6. Keseluruhan pengujian menunjukkan bahwa seleksi fitur berbasis probabilitas (IG, MI and CS) lebih baik dibanding seleksi fitur berbasis frekuensi (TF dan TFxIDF).

5. KESIMPULAN

Dalam makalah ini, telah dibangun sistem klasifikasi berita *hoax* menggunakan pendekatan statistik. Data yang ditetapkan untuk sistem berisi 220 artikel yang terdiri dari 89 artikel *hoax* dan 131 artikel non *hoax* yang diberi label secara manual. Sistem ini terdiri dari modul praproses, ekstraksi fitur, seleksi fitur dan klasifikasi itu sendiri. Untuk setiap modul, dilakukan eksperimen yang membandingkan beberapa teknik. Hasil eksperimen terbaik dicapai dengan algoritma *naïve bayes* dengan fitur unigram dimana seleksi fitur menggunakan operasi *union* antara *information gain* dan *mutual information*.

6. REFERENSI

Chen, Y. Y., Yong, S.-P., & Ishak, A. (2014): Email Hoax Detection System Using Levenshtein Distance Method. Journal of computers, vol. 9, no. 2, academy publisher.

Dimastyo, J. G., & Adisantoso, I. J. (2010): Pengukuran Kinerja Spam Filter dengan Feature Selection yang Berbeda Menggunakan Fungsi klasifikasi Multinomial Naïve Bayes, Makalah Kolokium.

Han, J., & Kamber, M. (2006): Data Mining Concept and Tehniques. San Fransisco: Morgan Kauffman. ISBN 13: 978-1-55860-901-3

Kotsiantis, Ikonomakis & Tampakas. (2005): Text Classification Using Machine Learning Techniques. WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, pp. 966-974.

Langgeni, D. P., Baizal, Z. A., & W., Y. F. (2010): Clustering artikel berita berbahasa indonesia menggunakan unsupervised feature selection.Seminar Nasional Informatika UPN "Veteran" Yogyakarta.

Manning, C. D., Raghavan, P., & Schutze, H. (2008): Introduction to Information Retrieval. America: Cambridge University Press.

Petkovic, T., Kostanj, Z., & Pale, P. (t.thn.): E-Mail System for Automatic Hoax Recognition.

R.Fanno. (1961): Transmission of Information. MIT Press, Cambridge, MA.

Sunjana. (2010): Aplikasi mining data mahasiswa dengan metode klasifikasi decision tree. Seminar Nasional Aplikasi Teknologi Informasi.

T.E. Dunning. (1993): Accurate Methods for Statistics of suprise and coincidence. In Computational Linguistic, volume 19:1, Hal 61-74.

Thorsten Joachims. (1997): A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In ICML-97.

Vuković, M., Pripužić, K., & Belani, H. (2009): An Intelligent Automatic Hoax Detection System. Knowledge-Based and Intelligent Information and Engineering Systems Lecture Notes in Computer Science Volume 5711, 318-325.

Wenyuan Dai, (1997): Transferring Naïve Bayes Classifiers for Text Classifications.

Yang, Y., & Liu, X., (1999): A Reexamination of text categorization methods. SIGIR-9.

Yang, Y., & Pedersen, J. O. (1997.): A Comparative Study on Feature Selection.