Walisongo Journal of Information Technology, Vol. 1 No. 1 (2019): 1-12
DOI : http://dx.doi.org/10.21580/wjit.2019.1.1.3915

Klasifikasi Berita Hoax Dengan Menggunakan Metode *Naive Bayes*

Hery Mustofa[1] Adzhal Arwani Mahfudh[2]
[1 2] Universitas Islam Negeri Walisongo Semarang
herymustofa@walisongo.ac.id, adzhal@walisongo.ac.id

Abstract

Hoaxes contain false news or non-sourced news. Today, hoaxes are very widely spread through internet media. The development of information technology that has so quickly triggered the spread of hoax information through the internet has become uncontrolled. So we need an intelligent system that can classify hoax news content that is spread through internet media. The hoax classification process can be done through the preprocessing stage then weighting the word and classification using naive bayes. Measurements were made using the 10-ford cross validation method. The results obtained from these measurements, it is known that the value of fold 6 has the highest accuracy, which is equal to 85.28% which is classified as relevant documents as much as 307 and irrelevant as much as 53 or an error rate of 14.72%. While the average value based on hoax news and true news value precision 0.896 and recall 0.853

Keywords : *hoax, klasifikasi, naive bayes, text minning*

Abstrak

Hoax dapat berarti berita bohong atau berita yang tidak mempunyai sumber. Saat ini, hoax sangat banyak tersebar melalui media internet. Perkembangan teknologi informasi yang begitu cepat memicu penyebaran informasi hoax melalui internet menjadi tidak terkontrol. Sehingga diperlukan suatu sistem cerdas yang dapat melakukan klasifikasi konten berita hoax yang tersebar melalu media internet. Proses klasifikasi hoax dapat dilakukan melalui tahap preprocessing kemudian pembobotan kata dan dilakukan klasifikasi menggunakan naive bayes. Pengukuran dilakukan dengan metode 10-ford cross validation. Dari pengukuran tersebut diperoleh hasil, nilai fold 6 mempunyai keakuratan tertinggi, yaitu sebesar 85.28 % yang mana dokumen terklasifikasi yang relevan sebanyak 307 dan dokumen tidak relevan sebanyak 53 atau error rate sebesar 14.72%. Sedangkan nilai rata-rata berdasarkan dokumen berita hoax dan dokumen berita benar nilai precision 0,896 dan recall 0.853.

Kata kunci : hoax, klasifikasi, naive bayes text minning

ISSN 2715-0143 (online) ISSN 2714-9048 (print)
http://journal.walisongo.ac.id/index.php/jit/index

1. PENDAHULUAN

Dalam Kamus Besar Bahasa Indonesia (KBBI) hoax mempunyai makna berita bohong, berita tidak bersumber. (Kemendikbud, 2019) Hoax adalah informasi sesat dan berbahaya. Hoax dapat menyesatkan persepsi atau pandangan manusia dengan menyampaikan atau menyebarkan informasi palsu sebagai suatu kebenaran. (Afriza & Adisantoso, 2018) Secara garis besar hoax adalah berita yang menyesatkan karena tidak mempunyai sumber yang bisa dipertanggungjawabkan dan bukti yang jelas. Berita hoax sengaja diciptakan oleh segelintir orang untuk memperoleh keuntungan pribadi demi tujuannya tercapai.

Saat ini berita hoax banyak tersebar melalui *internet*. Perkembangan teknologi informasi yang begitu cepat, memicu penyebaran informasi melalu *internet* yang tidak terkontrol, salah satu di dalamnya informasi dokumen yang mengandung hoax. Di Indonesia Teknologi informasi telah berkembang dengan sangat pesat di mana jumlah pengguna internet di Indonesia saat ini semakin terus bertambah. Menurut survei yang telah dilakukan oleh lembaga Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) pada tahun 2017 dijelaskan bahwa penetrasi pengguna internet Indonesia mencapai 143,26 juta jiwa atau 54,56% dari total populasi penduduk Indonesia 262 juta orang. (Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) , 2017) Untuk mengetahui berita informasi hoax atau fakta yang tersebar di internet, diperlukan metode klasifikasi dokumen secara manual maupun secara otomatis oleh sistem.

Klasifikasi dokumen yaitu proses atau metode dalam menemukan sekumpulan model yang mendeskripsikan dan membedakan kelas-kelas data sesuai dengan kategori yang dimilikinya. (Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) , 2017) Tujuan dari klasifikasi adalah untuk melakukan prediksi kelas dari obyek yang belum diketahui kelas dan karakteristik tipe datanya. Identifikasi konten hoax sudah dilakukan oleh komunitas internet yang tergabung di situs turnbackhoax.id. Situs tersebut dikelola oleh MAFINDO (masyarakat anti hoax Indonesia), sumber konten berasal dari laporan dari forum di jejaring sosial Facebook dengan nama FAFHH (forum anti fitnah hasut dan hoax). Metode identitas atau klasifikasi yang dilakukan pada situs turnbackhoax.id masih dilakukan secara manual, sehingga jika informasi semakin berkembang akan kesulitan dikarenakan informasi yang masuk semakin banyak. Dalam penelitian ini kan dilakukan klasifikasi menggunakan model pendekatan data mining sehingga klasifikasi dapat dilakukan oleh sistem secara otomatis.

Data mining adalah suatu proses ekstraksi atau penggalian data yang belum diketahui sebelumnya. (Connoly & Begg, 2015) Data tersebut digali informasinya berdasarkan database besar untuk diambil keputusan yang penting. Sedangkan klasifikasi adalah kumpulan model

yang dapat melakukan ilustrasi atau menggambarkan dan membedakan kelas data atau konsep, dengan tujuan mampu menggunakan model untuk melakukan prediksi kelas dari objek yang label kelasnya tidak diketahui. Model tersebut didasarkan pada pola analisis kumpulan data latih. (Han & Kamber, 2000)

Sebelum berita dilakukan klasifikasi maka diperlukan tahap *preprocessing*. *Preprocessing* mempunyai tahapan sebagai berikut, yaitu yang pertama *case folding,* kedua *tokenizing/parsing,* ketiga *filtering* dan ke empat *stemming*. Proses *stemming* menjadi tahapan paling penting di dalam tahap *preprocessing* dikarenakan pada *stemming* terjadi proses penghilangan kata imbuhan sehingga kata menjadi kata dasar. (Triawati, 2009) Model klasifikasi yang akan digunakan dalam penelitian ini adalah metode klasifikasi *naive bayes*. *Naive bayes* telah banyak digunakan pada data medis, jaringan komputer dan teks dikarenakan kesederhanaan, efektif dan kompabilitas menangkap visualisasi model data. (Abraham, 2009)

Penelitian terkait dengan klasifikasi konten berita pernah dilakukan sebelumnya, yaitu penelitian menggunakan metode *naive bayes* untuk klasifikasi berita olahraga dengan *enhanced confix stripping stemmer*. Terdapat 2 jenis dokumen berita yaitu berita latih dan berita uji. Berita latih didapat dari situs berita olahraga sport.detik.com. Peneliti mengambil 151 berita latih dari 6 kategori yaitu sepakbola, basket, raket, motoGP, formula 1 dan berita olahraga lainnya. Penelitian ini menggunakan datasets sebanyak 18 berita yang dipilih acak atau random. Dari 18 berita yang diujikan, diketahui terdapat 14 berita yang bernilai benar. Dari Penelitian itu dapat disimpulkan bahwa keakuratan klasifikasi *naive bayes* dengan *enhanced confix stripping stemmer* sebesar 77%. (Pramudita, Putro, & Makhmud, 2018)

Penelitian selanjutnya klasifikasi hoax pada dokumen berita bahasa indonesia yang mempunyai tema kesehatan dengan menggunakan metode *Modified K-Nearest Neighbor*. Proses klasifikasi hoax ini memiliki beberapa tahapan yaitu dimulai dari tahap *preprocessing*, pembobotan dan klasifikasi dengan menggunakan metode *Modified K-Nearsest Neighbor*. Datasets yang digunakan diperoleh dari jejaring sosial dan portal berita yang berjumlah 51 berita telah dilabeli oleh pakar secara manual dan 67 dilabeli oleh tim hoax *buster* secara manual serta 52 berita yang dilabeli oleh sistem secara otomatis. Dari penelitian itu diperoleh hasil sebagai berikut. Pengujian terbaik dengan nilai *k* berjumlah 4, dengan *precision* sebesar 0,83 *recall* sebesar 0,75, dan *f-measure* sebesar 0.79 serta menghasilkan akurasi akurasi sebesar 75%. (Prasetyo & Adikara, 2018)

Berdasarkan paparan di atas, diketahui bahwa klasifikasi *naive bayes* memberikan akurasi lebih tinggi dibandingkan dengan klasifikasi *K-Nearsest Neighbor*. Oleh karena itu, dalam penelitian ini akan dilakukan klasifikasi berita hoax menggunakan *naive bayes* . Tujuannya adalah untuk mengetahui tingkat akurasi metode

naive bayes digunakan untuk klasifikasi berita hoax berbahasa Indonesia.

2. METODE

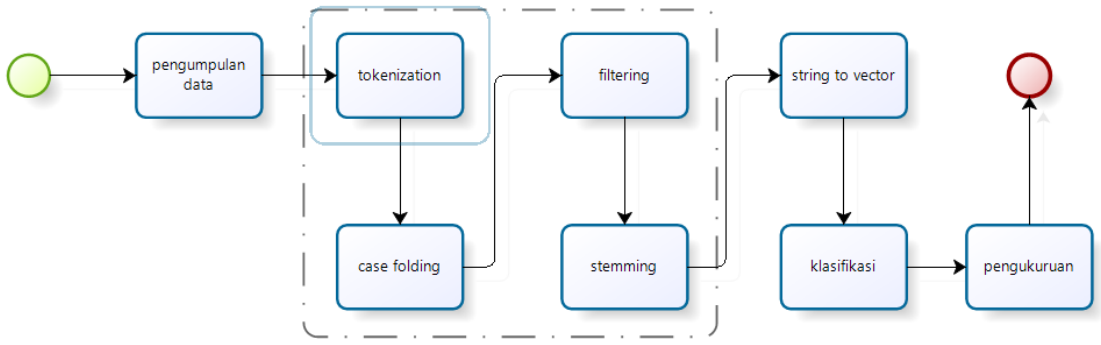
Metode penelitian ini akan menggunakan algoritma *naive bayes* dengan data masukan berupa dokumen teks. Setelah itu dokumen teks dilakukan *preprocessing* dengan tanpa menggunakan *steming*. Setelah itu, kemudian dilakukan proses pembobotan kata pada data latih (*data training*). Selanjutnya dilakukan klasifikasi teks berita tersebut dengan menggunakan algoritma *naive bayes*. Pengukuran dilakukan dengan menggunaka metode *10-Ford Cross Validation*. Hasil akhir dari proses klasifikasi menggunakan naive bayes akan menghasilkan *output* berupa berita yang sudah terlabeli hoax dan fakta. Tahapan sistem dapat terlihat seperti pada gambar 1.

3. KERANGKA TEORI

Di dalam kerangka teori akan dibahas dan dikupas beberapa hal atau istilah terkait dasar teori yang digunakan di dalam penelitian.

3.1 Berita

Berita dapat berupa informasi terbaru yang dapat menarik pembaca dan disampaikan lewat media seperti internet, koran, layar kaca, jejaring sosial, atau media lainya. Sebuah berita harus memuat unsur berikut, yaitu : *5W+1H* (*what, who, when, where, why dan How*). (Cahya, 2012)



Gambar 1
Metode Penelitian

3.2 Hoax

Berita yang mengandung makna berita bohong, berita tidak bersumber disebut hoax. (Kemendikbud, 2019) Informasi yang sesat dan berbahaya karena menyesatkan persepsi manusia dengan menyampaikan informasi palsu sebagai kebenaran disebut Hoax (Afriza & Adisantoso, 2018) Berita Hoax merupakan manipulasi berita yang sengaja dibuat dengan tujuan untuk memberikan informasi yang salah sehingga menyesatkan atau mengaburkan persepsi masyarakat.

3.3 Preprocessing

Preprocessing adalah proses melakukan perubahan dari dokumen teks menjadi *term index* dengan tujuan untuk menghasilkan *set term index* sehingga dapat digunakan sebagai *key word* untuk mengawali sebuah dokumen. (Fauzi, t.thn.) Tahapan *preprocessing* adalah sebagai berikut :

3.3.1 Parsing

Dokumen sebelum di proses, proses awal yaitu dilakukan *parsing*. Proses pemotongan struktur dokumen menjadi beberapa komponen yang terpisah disebut sebagai *parsing*.

3.3.2 Tokenisasi

Proses *tokenisasi* yaitu proses penghilangan angka, tanda baca, dan karakter selain huruf alfabet. Kemudian proses *tokeninsasi* selanjutnya yaitu pemotongan *string input* berdasarkan tiap kata penyusunya. Dikarenakan karakter tersebut merupakan suatu pemisah kata (*delimeter*) yang mana tidak memiliki kegunaan terhadap pemrosesan teks. Tahapan selanjutnya yaitu *case folding*. *Case folding* yaitu proses perubahan huruf besar menjadi kecil, di mana semua kata yang mengandung huruf besar diubah menjadi huruf kecil. Tahapan *case folding* terakhir yaitu tahapan *cleaning* mempunyai fungsi menghilangkan informasi yang tidak berhubungan dengan dokumen. Sebagai contoh yaitu *code script, link, HTML* dan lain sebagainya.

3.3.2 Filtering dan Stopword Removal

Di dalam tahapan ini melakukan proses *stoplist* atau *stopword.* Proses tersebut adalah melakukan penghilangan kata-kata yang tidak penting dengan pendekatan *bag-of-word*. Hasil dari *stoplist* adalah *wordlist* yang berisi kata penting.

3.3.3 Stemming

Stemming merupakan suatu teknik untuk mentransformasi kata-kata dalam sebuah dokumen teks menjadi bentuk kata dasar. Proses *stemming* berbeda-beda dalam tiap bahasa. Setiap bahasa memiliki aturan-aturan yang berbeda dalam penggunaan kata berimbuhan dan mempunyai aturan-aturan sendiri. Bahasa Perancis memiliki perbedaan aturan penggunaan tata bahasa dengan Bahasa Arab. Pada Bahasa Indonesia kopleksitas ada pada variasi imbuhan. Hal tersebut menjadi penting dalam pembentukan kata dasarnya.

3.4 Pembobotan Kata

Pembobotan kata yaitu penghitungan kata pada jumlah kemunculan masing-masing *token* dalam dokumen. Pembobotan kata yang paling banyak dipakai yaitu skema *term frequency-inverse document frequency* (TF-IDF). *Term frequency* (TF) didefinisikan sebagai jumlah kemunculan suatu kata/istilah dalam suatu dokumen. (Fauzi, t.thn.)

3.5 Term Frequency

Merupakan banyaknya tingkat kemunculan kata *t* dalam suatu

dokumen *d*. Rumus *term frequency* dapat dilakukan persamaan matematika sebagai berikut :

$$W_{t\,f\,t,d} = \begin{cases} 1 + log_{10} tf_{t,d,} & if\ tf_{t,d} > 0 \\ 0, & otherwise \end{cases}$$

Keterangan :
$W_{t\,f\,t,d} =$ Jumlah Frekuensi kemunculan kata *t* dalam dokumen *d*

3.6 Invers Document Frequency

Inverse document frequency atau *IDF* adalah banyak dokumen yang mengandung kata/*token*/*term t*. Rumus *IDF* dapat dinotasiakan sesuai pada persamaan berikut :

$$idf_t = \log_{10} \frac{N}{d\,f_t}$$

Keterangan :
idf_t = Banyaknya dokumen yang memuat t
N = Jumlah total dokumen

3.7 TF.IDF *Weighting*

TF.IDF Weighting adalah bobot hasil perkalian dari $t\,f_{\,t,d}$ dan idf_t, rumus *TF.IDF Weighting* dapat dinotasikan pada persamaan (3) dan normalisasinya dapata dinotasikan pada persamaan (4).

$$idf_t = \log_{10} \frac{N}{d\,f_t} \qquad (3)$$

$$W_{t,d} = \frac{W_{t\,d}}{\sqrt{\sum_t^n = 1\, W_{t,dd^2}}} \qquad (4)$$

3.8 Cousine Similarity

Merupakan proses untuk menghitung besarnya derajat kemiripan antara dokumen dan query. Nilai *cousine similarity* dapat dihitung berdasarkan perhitungan besarnya nilai fungsi *cosine* terhadap sudut yang dibentuk oleh dua buah vektor. Jika ditarik pada penelitian ini *cousine similarity* adalah sebuah representasi dari beberapa dokumen antar data latih. Rumus untuk menghitung tingkat kemiripan dokumen satu dengan dokumen lain dinotasikan pada persamaan (5) dan normalisasinya dapat dinotasikan pada persamaan (6).

Tanpa normalisasi *Wt, d* :

$$\boldsymbol{CosSim}(d_j, \boldsymbol{q}) = \frac{\boldsymbol{d_j}.\boldsymbol{q}}{|\boldsymbol{d_j}|.|\boldsymbol{q}|} = \frac{\sum_{i=1}^{t}(\boldsymbol{W_{ij}}.\boldsymbol{W_{iq}})}{\sqrt{\sum_{i=1}^{t} \boldsymbol{W_{ij}}^2.\sum_{i=1}^{t} \boldsymbol{W_{ij}}^2.\boldsymbol{W_{iq}}^2}}$$

Dengan normalisasi berdasarkan persamaan *Wt.d* sebelumnya :

$$\boldsymbol{CosSim}(d_j, \boldsymbol{q}) = d_j.q = \sum_{i=1}^{t}(W_{ij}.W_{iq}) \qquad (6)$$

Keterangan :
d_j = Data latih *j*
q = Tetangga data latih *j*
W_{ij} = Nilai pembobotan kata pada dokumen latih
W_{iq} = Nilai pembobotan kata pada tetangga dokumen latih

3.9 Naive Bayes

Bentuk analisis data yang mengekstrak model untuk menggambarkan kelas data dapat disebut sebagai klasifikasi. Sedangkan, model yang dibangun meliputi proses pengklasifikasian dan proses prediksi kategori berdasarkan *label* kelas. Klasifikasi dapat di aplikasikan ke dalam berbagai bidang seperti deteksi penipuan, target marketing, prediksi kerja dan lainnya.

Naive Bayes Classifier atau *NBC* merupakan proses pengklasifikasian probabilitas sederhana yang mengacu pada *Teory Bayes*. Teori tersebut menyatakan bahwa kemungkinan terjadinya suatu peristiwa sama dengan probabilitas intrinsik (dihitung dari data yang tersedia sekarang) dikalikan probabilitas bahwa hal serupa akan terjadi lagi di masa depan (berdasarkan pengetahuan yang terjadinya di masa lalu).

Naïve bayes adalah algoritma pembelajaran probabilitas yang berasal dari teori Keputusan Bayesian. Probabilitas *d* berada di kelas *c, P (c | d),* dihitung sebagai

$$P(c|d) \propto P\,(c) \prod_{k=1}^{m} P(t_k|\,c) \quad (7)$$

Di mana $P(t_k|\,c)$ adalah probabilitas bersyarat dari fitur t_k yang ada dalam pesan kelas c dan $P\,(c)$ adalah

probabilitas sebelumnya dari pesan yang terjadi di kelas *c*. $P(t_k|\,c)$

$$C_{MAP} = \arg \max_{cc\in\{cl\,,cs\,\}} P(c|d) = \arg \max_{c\in\{cl\,,cs\,\}} P\,(c) \prod_{k=1}^{m} P(t_k|\,c)$$

dapat digunakan untuk mengukur kontribusi *tk* ke dalam *c*, di mana c kelas yang benar. Dalam klasifikasi teks kelas pesan ditentukan dengan mencari kemungkinan besar atau maksimal posteriori *(MAP)* kelas C_{MAP} didefinisikan

Persamaan di bawah ini melibatkan banyak perkalian probabilitas bersyarat, salah satu fitur

dapat menghasilkan perhitungan titik bawah mengambang. Dalam prakteknya perkalian probabilitas di konversi ke dalam logaritma probabilitas oleh karena itu persamaanya didefinisakan

$$C_{MAP} = \arg \max_{c\in\{cl\,,cs\,\}} \left[\log P(c) + \sum_{k=1}^{m} logP(t_k|\,c)\right]$$

3.10 Pengukuran

Model klasifikasi yang telah dibangun perlu dilakuakan evaluasi atau pengukuran. Proses tersebut untuk mengetahui atau mengukur seberapa bagus model tersebut dalam melakukan klasifikasi yang di inginkan. Dalam melakukan evaluasi kinerja klasifikasi khususnya klasifikasi teks umumnya dilakukan dengan mengacu pada *accuracy* atau dengan *precision* and *recall* (Minner, Delen, & Elder, 2012). Nilai *accuracy*

merepresentasikan seberapa banyak keseluruhan dokumen yang dapat diklasifikasikan dengan benar. Semakin tinggi nilai *accuracy* yang dihasilkan maka semakin bagus dan akurat model tersebut dalam melakukan klasifikasi. Persamaan untuk mendapatkan nilai *accuracy* dapat dinotasikan sebagai berikut:

$$Acuracy = \frac{Total\ kata\ yang\ diklasifikasikan\ benar}{Total\ Dokumen}$$

$$P = \frac{TP}{(TP + FP)} \quad (10)$$

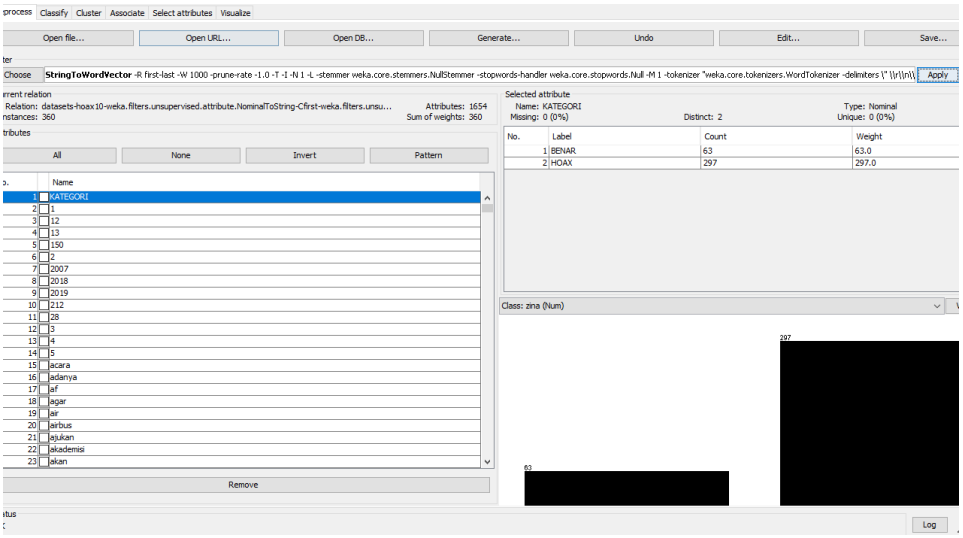
$$R = \frac{TP}{(TP + FN)} \quad (11)$$

Pada klasifikasi teks pengukuran *precision* dan *recall* merupakan metrik evaluasi yang paling banyak digunakan. Sebagai contoh: terdapat dua kelas A dan B, *precision* yaitu jumlah sampel berkategori A yang diprediksi dengan benar sebagai A dibanding dengan jumlah total data yang diprediksi sebagai A, sedangkan *recall* yaitu jumlah sampel berkategori A yang diprediksi dengan benar dibandingkan dengan jumlah total sampel A. Dalam melakukan pengukuran ini, biasanya dibangun table *confusion matrix* yang terdiri atas banyaknya baris data uji yang diprediksi benar dan tidak benar oleh model klasifikasi yang digunakan.

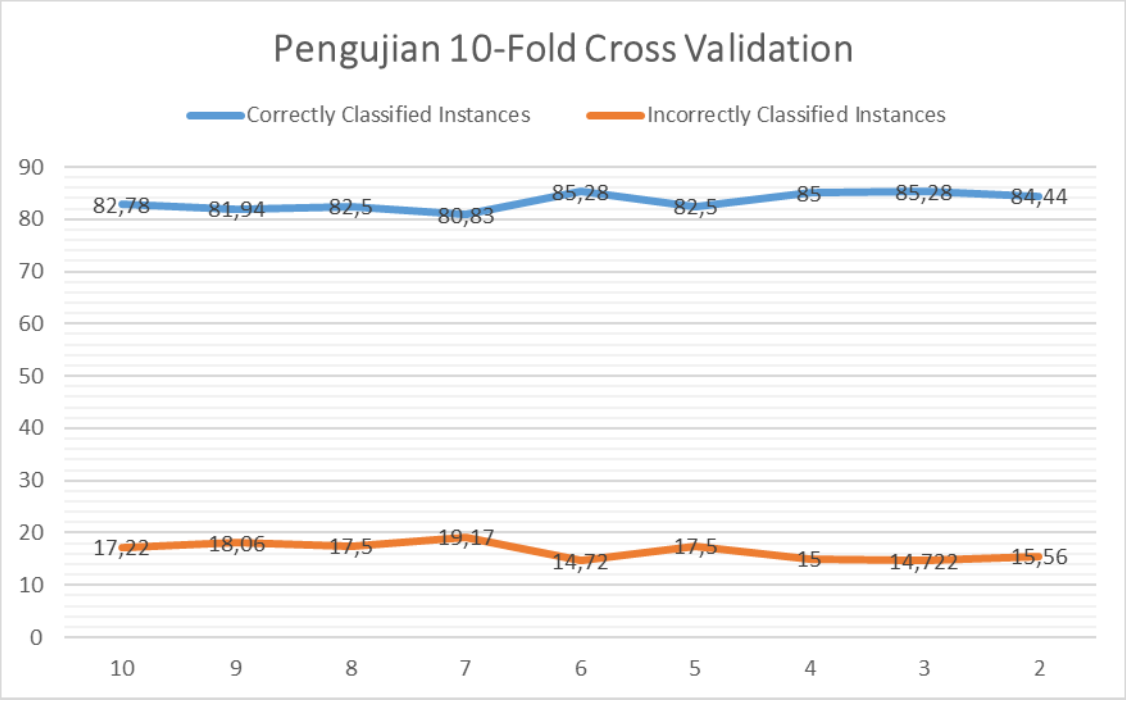
Tabel 1
Confuion Matrix

	Relevant	*Non relevant*
Retrieved	*True positives (TP)*	*False positives (FP)*
Not retrieved	*False negatives (FN)*	*Trus negatives (TN)*

Precision dan *recall* dapat merepresentasikan nilai keakuratan model pada setiap kelas. Sedangkan untuk mengetahui akurasi secara keseluruhan digunakan pengukuran *FI* yang merupakan pengukuran tunggal dari kombinasi *precision* dan *recall*.



Gambar 2
Proses Transformasi *String to Vector* dalam *Tool Weka*



Gambar 3
Pengujian *10-Fold Cross Validation*

4. PEMBAHASAN

Data berita hoax yang digunakan dalam penelitian ini diperoleh secara manual dari situs turnbackhoax.com. Data berita hoax diambil dari bulan November 2018 sampai dengan Februari 2019 berisi judul, tanggal berita terbit, dan label. Data yang diambil total sebanyak 300 di labeli oleh para pakar yang disebut sebagai data latih dan data tes. Dari 300 data tersebut terdapat 297 dokumen berita berkategori hoax, dan 63 dokumen berkategori fakta. Dokumen tersebut dikumpulkan dalam bentuk format *spreadsheet*.

Dari 300 data berita tersebut akan diolah menggunakan *Tool Weka 3.8.1*. Pertama yang dilakukan, yaitu melakukan tahap *preprocessing*. Tahap tersebut dimulai dari tahap *parsing* yaitu melakukan pemotongan dokumen, di mana data berformat *spreeadsheet* di rubah menjadi format *.arff* sesuai dengan format dokumen yang kompatibel *Tool Weka 3.8*. Setelah itu, dilakukan tahap tokenisasi. Tahap tersebut akan melakukan proses penghilangan tanda baca, spasi dan lain sebagainya. Selanjutnya dilakukan proses *stop word*, yaitu dokumen yang mengandung kata hoax dihilangkan, supaya tidak mengganggu hasil akurasi. Dari hasil *preprocessing* semua dokumen berita dapat terbentuk *attribut* sebanyak 1650 atribut. Seperti terlihat sesua pada gambar 2.

Tahapan selanjutnya, dilakukan proses transformasi kalimat ke dalam *matrix vector*. Dari proses transformasi tersebut, diketahui *TF* dan *IDF*. Setelah itu, dilakukan proses klasifikasi dengan *naive bayes*. Tahapan terakhir yaitu pengukuran dengan untuk menentukan nilai *accuracy, precision* and *recall* serta terbentuk *conffusion matrix*. Pengujian dalam penelitian ini menggunakan metode *10-Fold Cross Validation* dengan mengubah nilai *Fold* untuk dicari nilai akurasi terbaik. Hasil pengujian *10-Fold Cross Validation* sesuai terlihat sesuai pada gambar 3.

Hasil pengujian dengan metode *10-Fold Cross Validation*, diketahui bahwa nilai pengujian terbaik didapat dengan nilai *fold 6* dengan nilai keakuratan sebesar 85.28 % yang mana diketahui dokumen terklasifikasi yang relevan sebanyak 307 dan yang tidak relevan sebanyak 53 atau *error rate* sebesar 14.72%. Sedangkan nilai terendah didapat dari nilai *fold* 7 dengan nilai keakuratan 80.85% yang mana diketahui dokumen terklasifikasi yang relevan sebanyak 291 dan tidak relevan sebanyak 69 dokumen atau *error rate* sebesar 19,17%.

Nilai rata-rata berdasarkan dokumen berita hoax dan dokumen berita benar dengan metode pengujian *10-Fold Cross Validation* dengan nilai *fold 6* didapat nilai *precision* 0,896, *recall* 0.853 dan F-Measure 0.865. Hasil dapat terlihat pada Tabel *confusion matrix* tersaji dalam tabel 2.

Tabel 2
Confusion Matrix

Prediksi Hasil	Hasil Aktual	
	Fakta	H
Fakta	55	8
Hoax	45	252

5. PENUTUP

Metode *naive bayes* dapat digunakan pada sistem klasifikasi berita dengan masukan berupa teks dengan diawali tahap *preprocessing* yang berupa *parsing, tokenization, stopword*, dan pembobotan kata (*term weighting*). Kemudian dilakukan klasifikasi dengan metode *naive bayes*. Tahap terakhir yaitu dilakukan pengukuran dengan menggunakan pengujian *10-fold cross validation*. Dari hasil penelitian diketahui nilai *fold* 6 memberikan nilai akurasi dengan hasil terbaik dengan hasil dengan nilai keakuratan sebesar 85.28 % yang mana terklasifikasi dokumen yang relevan sebanyak 307 dan yang tidak relevan sebanyak 53 atau *error rate* sebesar 14.72%. Sedangkan nilai rata-rata berdasarkan berita hoax dan berita benar nilai *precision* 0,896 dan *recall* 0.853.

Dalam penentuan klasifikasi, sistem sangat bergantung dengan frekuensi kata dalam dokumen. Dalam penelitian selanjutnya jumlah data berita bisa ditambah dengan mencantumkan isi konten berita, jumlah kata semakin banyak akan mempengaruhi nilai akurasi.

Dalam penelitian ini, dikarenakan keterbatasan jumlah dokumen yang hanya berjumlah 360 dokumen, maka perlu dilakukan penelitian lanjutan dengan menambah total jumlah dokumen. Sehingga sistem semakin memiliki dataset yang beragam.

Dokumen yang diambil dalam penelitian ini masih bersifat umum, sehingga perlu dilakukan kajian penelitian dengan menggunakan konten dokumen yang bersifat khusus atau konten berita yang sebelumnya sudah terklasifikasi. Misalnya konten berita khusus kesehatan, konten berita khusus politik, konten berita khusus agama, dan sebagainya.

Selain itu, perlu dilakukan penelitian tentang *stemming* atau pencarian kata pada bentuk kata dasar khusus untuk bahasa Indonesia. Dengan dilakukan stemming akan mereduksi jumlah suku kata atau *attribut*. Dengan suku kata yang semakin berkurang akan mempengaruhi hasil klasifikasi konten berita hoax.

REFERENCES

Abraham, S. R. (2009). Effective Discretization and Hybrid Feature Selection Using Naive Bayesian Classifier For Medical Data Mining. *International Journal of Computational Intelligence Research 4*.

Afriza, A., & Adisantoso, J. (2018). Metode Klasifikasi Rocchio untuk Analisis Hoax. *Jurnal Ilmu Komputer Agri-Informatika, Volume 5 Nomor 1*, 1-10.

Asosiasi Penyelenggara Jasa Internet Indonesia (APJII) . (2017). *Infografis Penetrasi dan Pengguna Internet Indonesia.*

Cahya, I. (2012). *Menulis Berita di Media Massa.* Citra Aji Pratama.

Connoly, T. C., & Begg, C. E. (2015). *Database System: A Practical Approach to Design, Implementation, and Management.*

Fauzi, A. (t.thn.). *Text Mining 2017/2018*. Diambil kembali dari http://malifauzi.lecture.ub.ac.id/2017/09/text-mining-20172018/

Han, J. W., & Kamber, M. (2000). *Data Mining: Concepts and Techniques.*

Han, J., & Kamber, M. (2006). *Data Mining : Conceps and Techniques.* San Francisco: Elsevier Inc.

Kemendikbud, K. (2019, Juni 25). *Hasil Pencarian - KBBI Daring* . Diambil kembali dari https://kbbi.kemdikbud.go.id/entri/hoaks

Minner, G., Delen, D., & Elder, J. (2012). *Excerpt from: Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications.*

Pramudita, Y. D., Putro, S. S., & Makhmud, N. (2018). Klasifikasi Berita Olahraga Menggunakan Metode Naive Bayes dengan Enhanced Confix Stripping Stemmer. *Jurnal Teknologi Informasi dan Ilmu Komputer, Vol. 5, No. 3*, 269-276.

Severin, W. J., & James, J. T. (1998). Communication Theories: Origins, Methode, Uses (2th ed). New York: Longman Inc.

Ting, S. L., Ip, W. H., & Tsang, A. H. (2011). Is Naïve Bayes a Good Classifier for Document Classification? *International Journal of Software Engineering and Its Applications, 5*(3).

Triawati, C. (2009). Metode Pembobotan Statical Concept Based unuk Klastering dan Kategorisasi Dokumen Berbahasa Indonesia. *Institide Teknologi Telkom.* Bandung.