Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) DOI: 10.25126/jtiik.201854773
Vol. 5, No. 4, September 2018, hlm. 427-434 p-ISSN: 2355-7699
Akreditasi KEMENRISTEKDIKTI, No. 51/E/KPT/2017 e-ISSN: 2528-6579

PERBANDINGAN KINERJA METODE NAIVE BAYES DAN K-NEAREST NEIGHBOR UNTUK KLASIFIKASI ARTIKEL BERBAHASA INDONESIA

Riri Nada Devita[1], Heru Wahyu Herwanto[2], Aji Prasetya Wibawa[3]

[1]Teknik Informatika, Universitas Negeri Malang
[2,3]Dosen Teknik Elektro, Universitas Negeri Malang
Email: [1]ririnadadevita@gmail.com, [2]heru_wh@um.ac.id, [3]aji.prasetya.ft@um.ac.id

(Naskah masuk: 14 April 2018, diterima untuk diterbitkan: 23 September 2018)

Abstrak

Kecocokan isi artikel dengan sebuah tema jurnal menjadi faktor utama diterima tidaknya sebuah artikel. Tetapi masih banyak mahasiswa yang bingung untuk menentukan jurnal yang sesuai dengan artikel yang dimilikinya. Untuk itu diperlukannya sebuah metode klasifikasi dokumen yang dapat mengelompokkan artikel secara otomatis dan akurat. Terdapat banyak metode klasifikasi yang dapat digunakan. Metode yang digunakan dalam penelitian ini adalah *Naive Bayes* dan sebagai *baseline* digunakan metode *K-Nearest Neighbor*. Metode *Naive Bayes* dipilih karena dapat menghasilkan akurasi yang maksimal dengan data latih yang sedikit. Sedangkan metode *K-Nearest Neighbor* dipilih karena metode tersebut tangguh terhadap data *noise*. Kinerja dari kedua metode tersebut akan dibandingkan, sehingga dapat diketahui metode mana yang lebih baik dalam melakukan klasifikasi dokumen. Hasil yang didapatkan menunjukkan metode *Naive Bayes* memiliki kinerja yang lebih baik dengan tingkat akurasi 70%, sedangkan metode *K-Nearest Neighbor* memiliki tingkat akurasi yang cukup rendah yaitu 40%.

Kata kunci: *Klasifikasi dokumen, Naive Bayes, K-Nearest Neighbor*

Abstract

PERFORMANCE COMPARISON OF NAIVE BAYES AND K-NEAREST NEIGHBOR METHODS FOR INDONESIAN ARTICLES CLASSIFICATION

One way to be accepted in a journal conference and get the publication is to create an article with perfect suitability content of the journal. Matching the content of the article with a journal theme is the main factor for acceptability an article. But there are still many students who are confused to choose the journal in accordance with the articles it has. So we need a method to classification article documents category automatically and accurately group articles. There are many classification methods that can be used. The method used in this study is Naive Bayes and as a baseline the K-Nearest Neighbor method. Naive Bayes method is chosen because it can produce maximum accuracy with little training data. While K-Nearest Neighbor method was chosen because the method is robust to data noise. The performance of the two methods will be compared, so we can be known which method is better in classifying the document. The results show that the Naive Bayes method performs is more accurate with 70% accuracy and K-Nearest Neighbors method has a fairly low accuracy of 40% on classification test.

Keywords: *Documents classification, Naive Bayes, K-Nearest Neighbor*

1. PENDAHULUAN

Sebagai calon sarjana, mahasiswa tidak hanya menjadi konsumen ilmu pengetahuan. Seorang mahasiswa harus menjadi produsen yang mampu mengembangkan dan menerapkan ilmu pengetahuannya ke dalam sebuah karya ilmiah. Karya ilmiah sendiri merupakan suatu kegiatan menulis bagi mahasiswa untuk memaparkan hasil penelitian atau pengkajian yang telah dilakukannya sesuai metodologi penulisan yang baik dan benar (Lidwina, 2013). Karya tulis tersebut dipublikasi secara nasional maupun internasional.

Salah satu bentuk publikasi karya tulis ilmiah adalah jurnal. Untuk dapat diterima dalam jurnal dan mendapatkan publikasi, mahasiswa harus mempunyai karya tulis ilmiah dalam bentuk artikel yang memiliki kesesuaian isi dengan tema jurnal tersebut. Artikel yang memiliki keterkaitan dengan tema jurnal yang ditentukan akan direview terlebih dahulu sebelum dinyatakan diterima tidaknya artikel tersebut. Sedangkan isi artikel yang memiliki

keterkaitan tema terlalu jauh, akan langsung ditolak. Dengan demikian dapat disimpulkan bahwa kecocokan isi artikel dengan sebuah tema jurnal menjadi faktor utama diterima tidaknya sebuah artikel.

Pada prakteknya masih banyak mahasiswa yang bingung dalam menentukan jurnal yang sesuai dengan artikel yang dimilikinya, sehingga menyebabkan peluang diterimanya artikel sangat kecil. Salah satu cara untuk menentukan hal tersebut yaitu dengan membaca dan mencocokkan artikel yang dimilikinya dengan artikel yang telah terbit sebelumnya. Namun, cara seperti ini tentunya tidak efisien dan tidak akurat. Untuk itu diperlukannya sebuah metode klasifikasi dokumen yang dapat mengelompokkan artikel secara otomatis. Setiap jurnal pasti memiliki kata unik yang dapat merepresentasikan isi dari suatu jurnal. Hal tersebut yang digunakan sebagai acuan untuk melakukan klasifikasi secara akurat.

Terdapat banyak metode yang dapat digunakan untuk mengklasifikasikan dokumen, di antaranya yaitu *Support Vector Machine, K-Nearest Neighbor*, dan *Naive Bayes.* Penelitian dengan menggunakan metode *Naive Bayes* telah dilakukan oleh M. Ridwan untuk mengevaluasi kinerja akademik mahasiswa (Ridwan et al., 2013). Dari penelitian tersebut diketahui bahwa metode *Naive Bayes* berhasil melakukan klasifikasi dengan akurat. Metode tersebut hanya memerlukan data latih yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi (Saleh, 2015).

Penelitian selanjutnya dilakukan oleh Syahfitri dkk pada tahun 2015 untuk *Analysis Sentiment* pada Teks Bahasa Indonesia (Lidya et al., 2015). Hasil penelitian tersebut menunjukkan bahwa metode *K-Nearest Neighbor* mampu memberikan performa yang baik karena metode tersebut tangguh terhadap data *noise* (Bahri dan Maliki, 2012). Sedangkan metode *Support Vector Machine* memiliki tingkat akurasi yang baik tetapi memiliki waktu proses yang lama dibandingkan metode *K-Nearest Neighbor*.

Dikarenakan dataset yang akan digunakan dalam penelitian ini berskala kecil maka peneliti memutuskan untuk menggunakan metode *Naive Bayes* dan memilih metode *K-Nearest Neighbor* sebagai *baseline* untuk mengklasifikasikan artikel jurnal berbahasa Indonesia. Kinerja dari kedua metode tersebut akan dibandingkan, sehingga dapat diketahui metode yang paling efektif dalam melakukan klasifikasi artikel jurnal berbahasa Indonesia dengan jumlah dataset yang kecil.

Diharapkan dengan adanya penelitian ini dapat diketahui metode yang memiliki kinerja terbaik dalam melakukan klasifikasi dokumen berbahasa Indonesia. Metode dengan kinerja terbaik dapat digunakan untuk membangun sebuah sistem klasifikasi dokumen yang dapat memudahkan mahasiswa dalam memilih tema jurnal yang sesuai dengan artikel yang dimilikinya. Dengan sistem tersebut membuat peluang diterimanya suatu jurnal semakin besar sehingga mampu untuk mendapatkan publikasi.

2. PENELITIAN SEBELUMNYA

Terdapat banyak metode yang dapat digunakan dalam pengklasifikasian dokumen di antaranya yaitu *Support Vector Machine*, *K-Nearest Neighbor*, *Artificial Neural Network*, *Naive Bayes* dan masih banyak lagi. Penelitian dengan memanfaatkan metode *Support Vector Machine* dan *K-Nearest Neighbor* pernah dilakukan oleh Syahfitri dkk pada tahun 2015 dalam jurnalnya yang berjudul "Sentiment Analysis pada Teks Bahasa Indonesia menggunakan *Support Vector Machine* dan *K-Nearest Neighbor*" (Lidya et al., 2015). Pada penelitian tersebut dataset yang digunakan adalah data pemilu di Indonesia pada tahun 2014, dokumen hasil *crawling* terdiri dari 62545 *term*, yang diklasifikasikan menjadi 3 kelas yaitu positif, negatif, dan netral. Hasil percobaan dengan metode *K-Nearest Neighbor* menunjukkan bahwa metode ini mampu memberikan performa yang baik untuk data yang bersifat independen (tidak memiliki ketergantungan kata). Sedangkan untuk metode *Support Vector Machine* diketahui memiliki tingkat akurasi yang baik dan tidak dipengaruhi oleh besar kecilnya data uji.

Penelitian selanjutnya dilakukan oleh Rodrigo Moraes dkk, dalam jurnalnya yang berjudul "Document-level sentiment classification: An empirical comparison between SVM and ANN". Pada penelitian tersebut Rodrigo dkk melakukan komparasi metode klasifikasi SVM dengan Artificial Neural Network (ANN). Hasilnya, metode terbaik untuk klasifikasi adalah ANN. ANN mengungguli SVM dengan perbedaan yang signifikan secara statistik, bahkan pada konteks data yang tidak seimbang. Dilakukan 28 tes pada empat dataset, dan diketahui bahwa metode ANN mengungguli SVM secara signifikan) pada 13 tes, sementara SVM mengungguli ANN secara signifikan hanya dalam 2 tes.

Penelitian yang memanfaatkan metode klasifikasi berikutnya adalah "Seleksi Mobil Berdasarkan Fitur dengan Komparasi Metode Klasifikasi Neural Network, Support Vector Machine, dan metode C4.5", penelitian ini dilakukan oleh Purwaningsih pada tahun 2016 (Purwaningsih , 2016). Dalam penelitian tersebut diketahui bahwa metode C4.5 menghasilkan nilai AUC (Area Under ROC Curve) 0,888 dan model *Neural Network* dengan nilai AUC 0.884 sedangkan *Support Vector Machine* termasuk kategori *Fair Classification* dengan nilai AUC 0.793. Sehingga dapat ditaruk kesimpulan bahwa metode C4.5 memiliki tingkat akurasi yang paling tinggi dibandingkan dengan *Neural Network* dan *Support Vector Machine*. Hal ini dikarenakan terdapat kelemahan pada metode *Neural Network* di mana harus menggunakan data

pelatihan cukup besar untuk mendapatkan hasil yang maksimal, sedangkan kelemahan dari metode *Support Vector Machine* adalah sulit digunakan dalam jumlah sample berskala besar dan secara teoritik metode ini dikembangkan hanya untuk pengklasifikasian sebanyak dua *class* (Li et al., 2010).

Sedangkan penelitian dengan metode *Naive Bayes* dilakukan oleh M. Ridwan dkk untuk evaluasi kinerja akademik mahasiswa (Ridwan et al., 2013). Dalam penelitian tersebut mengungkapkan bahwa metode *Naive Bayes* berhasil melakukan prediksi dengan akurat, dan salah satu kelebihan dari *Naive Bayes* adalah tidak membutuhkan jumlah data latih yang besar untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian.

Dari uraian tersebut dapat disimpulkan bahwa setiap metode klasifikasi mempunyai kelebihan dan kelemahannya masing-masing. Selain itu metode tersebut dapat digunakan untuk membangun sebuah sistem atau aplikasi yang dapat membantu menyelesaikan berbagai masalah dikehidupan sehari-hari. Contohnya yaitu sistem untuk menganalisis berbagai komentar pada media sosial, aplikasi untuk mengevaluasi kinerja akademik dari mahasiswa, dan sistem untuk membantu menyeleksi dan memilih mobil dengan kualitas yang baik.

3. METODE PENELITIAN

3.1. Naive Bayes Classification (NBC)

Naive Bayes merupakan metode pengklasifikasian probabilistik sederhana. Metode ini akan menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Metode *naive bayes* menganggap semua atribut pada setiap kategori tidak memiliki ketergantungan satu sama lain (*independen*) (Nafalski & Wibawa, 2016). Keuntungan penggunaan *Naive Bayes* yaitu hanya memerlukan sejumlah kecil data latih untuk menentukan parameter *mean* dan *varians* dari variabel yang diperlukan untuk klasifikasi (Palaniappan dan Awang, 2008). *Naive Bayes* merupakan metode *supervised document classification* yang berarti membutuhkan data training sebelum melakukan proses klasifikasi.

Dalam proses pelatihan, dokumen telah ditentukan kategorinya (data latih), yang kemudian akan diproses dan membentuk pengetahuan berupa nilai probabilitas pada setiap kata. Proses ini akan menghasilkan sebuah kata pada setiap dokumen yang mengkarakteristikkan dokumen pada suatu kategori tertentu. Untuk menghitung setiap kata yang terdapat pada dokumen latih dapat digunakan Persamaan 1, sedangkan untuk menghitung probabilitas kategori dokumen digunakan Persamaan 2.

$$p(w_i\,|c_j) = \frac{1+ n_i}{n+ |kosakata|} \quad (1)$$

dimana:

$p(w_i\,|c_j)$: probabilitas kata pada setiap kategori
n_i : frekuensi kemunculan kata setiap kategori
n : jumlah seluruh kata dalam dokumen pada kategori tertentu
$|kosakata|$: jumlah total kata di semua data latih

$$p(c_j) = \frac{n(doc_j)}{n\ (sampel)} \quad (2)$$

dimana:

$p(c_j)$: probabilitas dokumen kategori
$n(doc_j)$: jumlah seluruh dokumen pada suatu kategori
$n\ (sampel)$: jumlah seluruh dokumen latih

Setelah melakukan proses pelatihan, selanjutnya yaitu proses klasifikasi. Pada proses ini dokumen yang digunakan belum diketahui kategorinya (data uji), sehingga metode *naive bayes* akan mencari kata pada data uji yang sesuai dengan pengetahuan di data latih $p(w_i|c_j)$. Kemudian hitung probabilitas setiap dokumen $p(c_j)$ yang telah disimpan di pengetahuan pada saat proses pelatihan sebelumnya, maka untuk setiap kategori dokumen dapat dihitung menggunakan Persamaan 3.

$$p(c_j) \prod_i p(w_i|c_j) \quad (3)$$

Selanjutnya untuk mencari nilai $p(w_i|c_j)$ dapat dilakukan dengan cara mengalikan nilai probabilitas kemunculan kata yang sama pada data latih dengan nilai probabilitas dokumen yang sesuai kategorinya $p(c_j)$. Setelah didapatkan hasil perkalian pada masing-masing kategori dokumen, selanjutnya yaitu membandingkan dan mencari nilai probabilitas terbesar c_{MAP} yang digunakan untuk klasifikasi data uji pada dokumen jurnal bahasa indonesia yang akan diklasifikasikan ke dalam salah satu kategori yang tersedia (Schneider, 2005), perhitungan tersebut dapat dilihat pada Persamaan 4.

$$c_{MAP} = argmax_{c_j \in c}\, p(c_j) \prod_i p(w_i|c_j) \quad (4).$$

3.2. K-Nearest Neighbor

K-Nearest Neighbor (KNN) adalah sebuah metode *supervised* yang berarti membutuhkan data *training* untuk mengklasifikasikan objek yang jaraknya paling dekat. Prinsip kerja *K-Nearest Neighbor* adalah mencari jarak terdekat antara data yang akan di evaluasi dengan *k* tetangga (*neighbor*) dalam data pelatihan (Whidhiasih et al., 2013).

Pada proses pelatihan, dokumen dikelompokkan secara manual sesuai dengan kategori yang telah ditentukan. Setelah itu dokumen tersebut akan melalui tahapan *preprocessing* yang akan menghasilkan bobot untuk setiap kata yang ada di semua dokumen latih. Selanjutnya

menghitung kemiripan vektor dokumen uji dengan setiap dokumen latih yang telah di klasifikasikan. Untuk mengetahui kemiripan dokumen digunakan metode *cosine similarity* (Ridok dan Indriati, 2015). Metode ini dapat digunakan untuk menginterpretasikan jarak tiap dokumen berdasarkan kemiripan dokumen (Rivki dan Bachtiar, 2017). Perhitungan jarak dengan metode *cosine similarity* dapat dilihat pada Persamaan 5.

$$Cos(i,k) = \frac{\sum_k (d_i d_k)}{\sqrt{\sum_k d_{ik}^2}\sqrt{\sum_k d_{jk}^2}} \quad (5)$$

dimana:

$\sum_k (d_i d_k)$: *vector* dari produk *i* dan *k*
$\sqrt{\sum_k d_{ik}^2}$: panjang dari *vector i*
$\sqrt{\sum_k d_{jk}^2}$: panjang dari *vector j*
i : data uji ke-*i*
j : data latih ke-*j*

Selanjutnya yaitu mengurutkan jarak tersebut berdasarkan nilai terkecil (terdekat) hingga yang terbesar (terjauh). Kemudian menentukan jumlah tetangga (nilai *k*) yang ingin digunakan sebagai acuan untuk proses klasifikasi. Dari nilai *k* inilah dapat ditentukan kategori dokumen berdasarkan nilai *euclidean* terdekat.

3.3. Pembobotan Kata (TF-IDF)

Term frequence (TF) dan *Inverse document Frequency* (IDF) adalah pembobotan yang paling sering digunakan (Spärck Jones, 2004). Metode TF-IDF merupakan cara untuk mencari bobot suatu kata (*term*) pada sebuah dokumen (Robertson, 2004). Metode TF-IDF menggabungkan dua cara untuk perhitungan bobotnya, yaitu dengan menghitung frekuensi kemunculan kata di sebuah dokumen tertentu (TF) dan melakukan perhitungan *invers* terhadap frekuensi dokumen yang mengandung kata tersebut (IDF) (Prabowo et al., 2016). Perhitungan *invers document frequency* (IDF) digunakan untuk menghitung kuantitas *term* yang berfungsi sebagai ukuran tingkat signifikansi suatu *term* dalam sebuah dokumen (Pujianto, 2013). Perhitungan TF dan IDF dapat dilihat pada Persamaan 6 dan 7 (Fauzi et al., 2015).

$$TF(d,t) = f(d,t) \quad (6)$$

$$IDF(t) = 1 + log\left(\frac{Nd}{df(t)}\right) \quad (7)$$

dimana:

$f(d,t)$: frekuensi kemunculan *term t* pada dokumen *d*
Nd : jumlah seluruh dokumen
$df(t)$: jumlah dokumen yang terdapat *term t*

Sehingga untuk menemukan nilai TF-IDF dapat digunakan Persamaan 8.

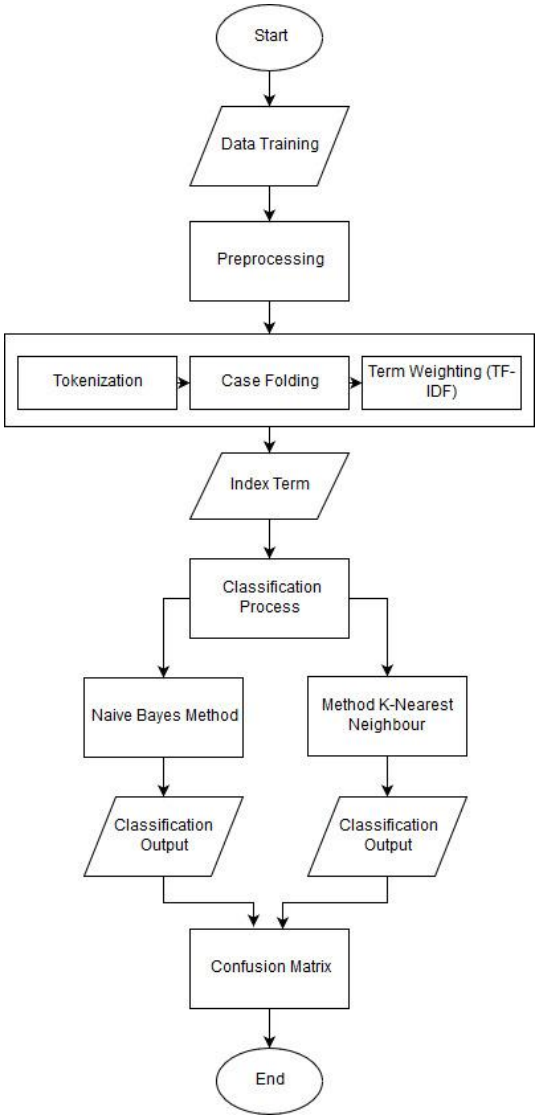
$$TF - IDF = TF(d,t).IDF(t) \quad (8)$$

3.4. Text Preprocessing

Text preprocessing bertujuan untuk mempersiapkan dokumen teks menjadi data siap diolah pada proses selanjutnya. Adapun tahapan *preprocessing* yang dilakukan, yaitu:

- *Tokenizing*
  Tahapan untuk melakukan pemenggalan kata pada dokumen. Spasi digunakan sebagai pemisah antar katanya. Pada tahap ini juga akan dilakukan pemfilteran dengan membuang karakter tertentu, seperti tanda baca.
- *Case folding*
  Tahapan untuk mengubah semua karakter huruf kapital di dalam dokumen menjadi huruf kecil. Karakter selain huruf a-z akan dianggap sebagai *delimiter*.

Untuk lebih jelasnya alur kerja dari klasifikasi dokumen dapat dilihat pada Gambar 1.



Gambar 1. Alur Klasifikasi Dokumen

3.5. Metode Pengumpulan Data

Teknik pengumpulan data dan informasi yang digunakan dalam penelitian ini adalah dengan melakukan studi literatur yang akan menghasilkan data sekunder. Data yang digunakan untuk klasifikasi adalah abstrak dari artikel jurnal berbahasa Indonesia yang ada di Universitas Negeri Malang. Dokumen yang akan digunakan sebanyak empat puluh dokumen jurnal berbahasa Indonesia. Dokumen tersebut diunduh dari website *journal2.um.ac.id*. Atribut yang akan digunakan dalam penelitian ini dapat dilihat pada Tabel 1.

Tabel 1. Atribut Pada Data Uji

Nama Atribut	Keterangan	Tipe Data
Text	Abstrak dokumen jurnal yang akan diklasfiikasikan	String
Class-at	Kategori Jurnal	Nominal

3.6. Metode Pengujian

Evaluasi dari hasil klasifikasi dokumen jurnal berbahasa Indonesia dilakukan dengan *confusion matrix*. Metode ini merepresentasikan hasil klasifikasi menggunakan matriks yang dapat dilihat pada Tabel 2.

Tabel 2. Confusion Matrix

		Clasified as	
Correct Classification		+	-
	+	True positive	False positive
	-	False negative	True negative

True Positive adalah jumlah *record* positif yang berhasil diklasifikasikan sebagai positif, sedangkan *false positive* merupakan *record* positif yang salah diklasifikasikan menjadi negatif. Sedangkan *false negative* merupakan *record* negatif yang salah diklasifikasikan sebagai positif, dan untuk *true negative* adalah *record* negatif yang berhasil diklasifikasikan sebagai *record* negatif. Metode pengujian *confusion matrix* dapat menghasilkan perhitungan dengan 4 output, di antaranya yaitu:

$$Precision = \frac{TP}{TP+FP}\ x\ 100\% \quad (9)$$

$$Recall = \frac{TP}{TP+FN}\ x\ 100\% \quad (10)$$

$$Akurasi = \frac{TP+TN}{TP+TN+FP\_FN}\ x\ 100\% \quad (11)$$

$$Error\ = \frac{FP+FN}{TP+TN+FP+FN}\ x\ 100\% \quad (12)$$

4. HASIL DAN PEMBAHASAN

4.1. Dataset

Dataset yang digunakan sebanyak 40 jurnal yang telah dipublikasi 2 tahun terakhir, jurnal tersebut di antaranya yaitu; jurnal Pendidikan Ekonomi, Pendidikan Bisnis dan Manajemen, Akuntansi Aktual, dan jurnal Ekonomi Bisnis. Masing-masing jurnal diambil sebanyak 10 jurnal.

4.2. Hasil *Preprocessing*

Preprocessing dilakukan agar data dapat diproses ke tahap klasifikasi. Adapun tahapan *preprocessing* yang dapat dilihat pada Gambar 1. Beberapa hasil *tokenisasi* dan perhitungan TF-IDF pada masing-masing kategori dokumen dapat dilihat pada Tabel 3. Dimana *Dok A* adalah dokumen dengan kategori Pendidikan Ekonomi, *Dok B* adalah kategori Pendidikan Bisnis dan Manajemen, *Dok C* adalah kategori Akuntansi Aktual, dan *Dok D* adalah dokumen dengan kategori Ekonomi Bisnis.

Tabel 3. Hasil Tokenisasi dan TF-IDF

	TF-IDF			
Term	**Dok A**	**Dok B**	**Dock C**	**Dok D**
anggaran	0	0	1,795	0

asosiatif	0	0	0	2,556
dividen	0	0	0	1,596
efikasi	0	2,556	0	0
equity	0	0	0	1,596
hermeneutik	0	0	2,556	0
Kualitatif	0	0	1,033	0
kuantitatif	0	0	0	1,115
Metode	0	0,516	0,51	0,516
observasi	0	0	1,208	0
passing	2,556	0	0	0
perspektif	0	0	2,076	0
Populasi	0	0,834	0	0,834
proportionate	0	2,076	0	0
purposive	0	0	0	1,441
ratio	0	0	0	1,596
regresi	0	0,894	0	0
signifikan	0	0	0	0,894
Sampel	0	0,679	0	0
variabel	0	1,208	0	0

4.3. Hasil Klasifikasi

Setelah melakukan *preprocessing* selanjutnya dilakukan tahap klasifikasi dengan menggunakan metode *naive bayes* dan *k-nearest neighbor*. Proses klasifikasi bertujuan untuk mengklasifikasikan suatu data ke dalam kelompok kelas yang sudah ada yaitu Pendidikan Ekonomi, Pendidikan Bisnis & Manajemen, Akuntansi Aktual dan Ekonomi Bisnis. Pengukuran kinerja metode klasifikasi akan dilakukan dengan *k-fold cross validation method* yang akan menghasilkan *confusion matrix* yang dapat dilihat pada Tabel 6 dan 7.

Teknik pengujian menggunakan *k-fold cross validation* ini sangat cocok digunakan untuk mengolah data yang jumlahnya sedikit. Prinsip kerjanya yaitu, membagi data sebanyak *k* sub-himpunan, dimana *k* adalah nilai dari *fold*. Selanjutnya tiap sub-himpunan tersebut akan dijadikan data uji dari hasil klasifikasi yang dihasilkan dari *k-1* sub-himpunan lainnya. Setiap datum akan menjadi data uji sebanyak 1 kali, dan menjadi data training sebanyak *k-1* kali.

Penelitian ini menggunakan nilai *fold=10,* sehingga dari 40 data akan dibagi menjadi 10 blok dengan jumlah training yang sama yaitu 4 *instance*. Setiap datum akan menjadi data testing 1 kali dan menjadi data training sebanyak 3 kali (*k-1)*. Beberapa dokumen hasil klasifikasi dapat dilihat pada Tabel 4 untuk metode *Naive Bayes* dan Tabel 5 untuk metode *K-Nearest Neighbor*.

Tabel 4. Klasifikasi Naive Bayes

Dok ke-	Klasifikasi Sebenarnya	Hasil Klasifikasi	Ket
39	Ekonomi Bisnis	Ekonomi Bisnis	TRUE
2	Pendidikan Ekonoomi	Pendidikan Ekonomi	TRUE
28	Akuntansi Aktual	Ekonomi Bisnis	FALSE
13	Pendidikan Bisnis Manajemen	Pendidikan Bisnis Manajemen	TRUE
34	Ekonomi Bisnis	Ekonomi Bisnis	TRUE
4	Pendidikan Ekonoomi	Pendidikan Ekonomi	TRUE
22	Akuntansi Aktual	Ekonomi Bisnis	FALSE
15	Pendidikan Bisnis Manajemen	Pendidikan Ekonomi	FALSE
36	Ekonomi Bisnis	Ekonomi Bisnis	TRUE
7	Pendidikan Ekonoomi	Ekonomi Bisnis	FALSE

Tabel 5. Klasifikasi K-Nearest Neighbor

Dok ke-	Klasifikasi Sebenarnya	Hasil Klasifikasi	Ket
39	Ekonomi Bisnis	Pendidikan Bisnis Manajemen	FALSE
2	Pendidikan Ekonoomi	Pendidikan Bisnis Manajemen	FALSE
28	Akuntansi Aktual	Pendidikan Bisnis Manajemen	FALSE
13	Pendidikan Bisnis Manajemen	Pendidikan Bisnis Manajemen	TRUE
34	Ekonomi Bisnis	Ekonomi Bisnis	TRUE
4	Pendidikan Ekonoomi	Pendidikan Ekonomi	TRUE
22	Akuntansi Aktual	Pendidikan Bisnis Manajemen	FALSE
15	Pendidikan Bisnis Manajemen	Pendidikan Bisnis Manajemen	TRUE
36	Ekonomi Bisnis	Ekonomi Bisnis	TRUE
7	Pendidikan Ekonoomi	Pendidikan Bisnis Manajemen	FALSE

Data akan bernilai *True* jika kategori dokumen hasil klasifikasi menggunakan metode *naive bayes* atau *KNN* sama dengan kategori data sebenarnya dan akan bernilai *False* apabila kelas hasil klasifikasi tidak sama dengan kelas sebenarnya. *Confusion matrix* dari masing-masing metode dapat dilihat pada Tabel 6 dan 7.

Tabel 6. Confusion Matrix Naive Bayes

a	b	c	d	Classified as
8	1	0	1	a= PendidikanEkonomi
5	2	1	2	b= PendidikanBisnisManajemen
0	0	8	2	c= AkuntansiAktual
0	0	0	10	d= EkonomiBisnis

Tabel 6 menunjukkan bahwa, *a* adalah jurnal dengan kategori *PendidikanEkonomi*, *b* jurnal *PendidikanBisnisManajemen*, *c* jurnal *Akuntansi-Aktual*, sedangkan *d* merupakan jurnal dengan kategori *EkonomiBisnis*. Diketahui bahwa terdapat 10 dokumen dengan kategori jurnal *Pendidikan-Ekonomi*, dari 10 dokumen tersebut metode *naive bayes* berhasil mengklasifikasikan 8 dokumen sesuai kelasnya, sedangkan 2 lainnya salah diklasifikasikan sebagai jurnal *PendidikanBisnisManajemen* dan jurnal *EkonomiBisnis.*

Untuk jurnal kategori *PendidikanBisnis-Manajemen* yang berjumlah 10 dokumen, 2 dokumen berhasil diklasifikasikan sesuai kelasnya, sedangkan 5 dokumen lainnya salah diklasifikasikan sebagai jurnal *PendidikanEkonomi*, 1 dokumen sebagai jurnal *AkuntansiAktual*, dan 2 dokumen diklasifikasikan sebagai jurnal *EkonomiBisnis.*

Selanjutnya terdapat jurnal kategori *AkuntansiAktual*. Dari 10 dokumen terdapat 8 dokumen yang diklasifikasikan sesuai dengan

kelasnya, sedangkan 2 dokumen lainnya salah diklasifikasikan sebagai jurnal *EkonomiBisnis*. Kemudian terdapat jurnal *EkonomiBisnis* sebanyak 10 dokumen, dan semua jurnal tersebut terklasifikasi sesuai kelasnya.

Selanjutnya terdapat *confusion matrix* dari metode *K-Nearest Neighbor* yang dapat dilihat pada Tabel 7. Pada penelitian ini penulis menetapkan jumlah dokumen tetangga sebanyak k=5.

Tabel 7. Confusion Matrix K-Nearest Neighbor

a	b	c	d	Classified as
3	7	0	0	a= PendidikanEkonomi
3	6	0	1	b= PendidikanBisnisManajemen
0	8	2	0	c= AkuntansiAktual
0	5	0	5	d= EkonomiBisnis

Confusion matrix dari metode *K-Nearest Neighbor* menunjukkan bahwa, dari 10 dokumen dengan kategori jurnal *PendidikanEkonomi*, terdapat 3 dokumen yang diklasifikasikan sesuai kelasnya, sedangkan 7 dokumen lainnya salah diklasifikasikan menjadi jurnal *PendidikanBisnisManajemen*. Untuk jurnal kategori *PendidikanBisnisManajemen* yang berjumlah 10 dokumen, terdapat 6 dokumen yang diklasifikasikan sesuai dengan kelasnya, sedangkan 3 dokumen salah diklasifikasikan menjadi jurnal *PendidikanEkonomi* dan 1 dokumen lainnya salah diklasifikasikan menjadi jurnal *EkonomiBisnis*. Selanjutnya dokumen dengan kategori jurnal *AkuntansiAktual*, dari 10 dokumen terdapat 2 dokumen yang diklasifikasikan sesuai dengan kelasnya, sedangkan 8 dokumen lainnya salah diklasifikasikan sebagai jurnal *AkuntansiAktual*. Selanjutnya untuk kategori jurnal *EkonomiBisnis*, metode *KNN* berhasil mengklasifikasikan 5 dari 10 dokumen sesuai dengan kelasnya, sedangkan 5 dokumen lainnya salah diklasifikasikan sebagai jurnal *PendidikanBisnisManajemen.*

4.4. Hasil Pengujian

Berdasarkan 40 dokumen yang telah diuji, didapatkan hasil perhitungan *precision*, *recall*, *accuracy* dan *error* dari masing-masing metode. Hasil pengujian setiap metode dapat dilihat pada Tabel 8.

Tabel 8. Perbandingan Kinerja

Metode	Accuracy	Recall	Precision	Error
Naive Bayes	70%	70%	70,9%	30%
K-Nearest Neighbor	40%	40%	64,1%	60%

Berdasarkan Tabel 8 dapat diketahui bahwa kinerja dari metode *Naive bayes* lebih baik dari metode *K-Nearest Neighbor*. Bagaimanapun akurasi klasifikasi tidak dapat mencapai hasil yang sempurna dengan tidak adanya *error*. Hal tersebut dipengaruhi oleh banyaknya data uji dan data latih yang digunakan dan tahapan *preprocessing* yang dilakukan.

Untuk algortima *naive bayes* akurasi yang dihasilkan cukup baik, hal ini karena keunggulan dari metode *naive bayes* sendiri yaitu mampu melakukan klasifikasi meskipun memiliki data *training* yang sedikit untuk estimasi parameternya. Sedangkan untuk metode *K-Nearest Neighbor* menghasilkan akurasi yang rendah, hal ini dikarenakan metode tersebut tidak efektif jika data latih jumlahnya sedikit.

5. KESIMPULAN

Setelah menerapkan metode *Naive Bayes* dan *K-Nearest Neighbor* untuk mengklasifikasikan artikel jurnal berbahasa Indonesia diketahui bahwa kinerja dari metode *Naive Bayes* lebih unggul dari metode *K-Nearest Neighbor*. Terbukti bahwa dari 40 data uji yang digunakan metode *Naive Bayes* mampu mengklasifikasikan artikel jurnal berbahasa Indonesia sebanyak 28 dokumen. Sedangkan untuk metode *K-Nearest Neighbor* dari 40 data uji metode ini hanya dapat mengklasifikasikan artikel jurnal berbahasa Indonesia sebanyak 16 dokumen. Hal tersebut dapat dipengaruhi jumlah data yang digunakan dan tahapan *preprocessing* yang dilakukan. Oleh karena itu, disarankan untuk menambah data set dan melengkapi tahapan *preprocessing* seperti melakukan *stemming* kata pada penelitian selanjutnya.

6. DAFTAR PUSTAKA

BAHRI, R.S. & MALIKI, I., 2012. Perbandingan Algoritma Template Matching dan Feature Extraction pada Optical Character Recognition. *Jurnal Komputer dan Informatika (Komputa)*, 1(1), Pp.187–198. Available At: Http://Repo.Pens.Ac.Id/1324/1/Paper_Ta_Mbah.Pdf.

FAUZI, M. A., ARIFIN, A.Z., YUNIARTI, A., 2015. Term Weighting Berbasis Indeks Buku dan Kelas untuk Perangkingan Dokumen Berbahasa Arab. *Lontar Komputer*, 5(2), Pp.110–117.

LI, M., WANG, M. & WANG, C., 2010. Research On Svm Classification Performance In Rolling Bearing Diagnosis. In *2010 International Conference On Intelligent Computation Technology And Automation, ICICTA 2010*. Pp. 132–135.

LIDWINA, S., 2013. Penulisan Paragraf dalam Karya Ilmiah Mahasiswa. *Jurnal STIE Semarang*, 5(1), Pp.38–47.

LIDYA, S.K., SITOMPUL, O.S. & EFENDI, S., 2015. Sentiment Analysis pada Teks Bahasa Indonesia menggunakan Support Vector Machine (SVM) dan K-Nearest Neighbor (K-NN). *Seminar Nasional Teknologi dan Komunikasi* (SENTIKA), Pp.1–8.

NAFALSKI, A., & WIBAWA, A. P. 2016. Machine Translation With Javanese Speech Levels' Classification. Informatics, Control, Measurement in Economy and Environment Protection, 6(1), 21–25. https://doi.org/ 10.5604/20830157. 1194260

NURGIYANTORO, B., 2004. Penilaian Pembelajaran Sastra Berbasis Kompetensi. *DIKSI*, 11(1), Pp.91–116.

PALANIAPPAN, S. & AWANG, R., 2008. Intelligent Heart Disease Prediction System Using Data Mining Techniques. In *2008 IEEE/ACS International Conference On Computer Systems And Applications*. Pp. 108–115. Available At: Http://Ieeexplore.Ieee.Org/Lpdocs/Epic03/Wrapper.Htm?Arnumber=4493524.

PRABOWO, D. A., FHADLI, M., NAJIB, M. A., & FAUZI, H.A., 2016. Tf-Idf- Enhanced Genetic Algorithm Untuk Extractive Automatic Text Summarization. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, 3(3), pp.208–215.

PUJIANTO, U. 2013. Using Cosine Similarity For Determining The Inversed Document Frequency Value of Newly Added Documents. In Seminar On Electrical, Informatics, and its Education 2013, pp. 141–144.

PURWANINGSIH, E., 2016. Seleksi Mobil Berdasarkan Fitur dengan Komparasi Metode Klasifikasi Neural Network, Support Vector Machine, dan Algoritma C4.5. *Jurnal Pilar Nusa Mandiri*, XII(2), Pp.153–160.

RIDOK, A. & INDRIATI, 2015. Pengklasifikasian Dokumen Berbahasa Indonesia dengan Pengindeksan Berbasis LSI. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, 2(2), Pp.87–95.

RIDWAN, M., SUYONO, H. & SAROSA, M., 2013. Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa menggunakan Algoritma Naive Bayes Classifier. *EECCIS*, 7(1), Pp.59–64.

RIVKI, M. & BACHTIAR, A.M., 2017. Implementasi Algoritma K-Nearest Neighbor dalam Pengklasifikasian Follower Twitter yang Menggunakan Bahasa Indonesia. *Jurnal Sistem Informasi*, 13(1), Pp.31–37.

ROBERTSON, S., 2004. Understanding Inverse Document Frequency: On Theoretical Arguments For Idf. *Journal Of Documentation*, 60(5), Pp.503–520. Available At: Http://Www.Emeraldinsight.Com/Doi/10.1108/00220410410560582.

SALEH, A., 2015. Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. *Creative Information Technology Journal*. 2. 207-217.

SCHNEIDER, K.M., 2005. Techniques For Improving The Performance Of Naive Bayes for Text Classification. in *Computational Linguistics and Intelligent Text Processing*. Pp. 682–693.

SPÄRCK JONES, K., 2004. A Statistical Interpretation of Term Specificity and Its Retrieval. *Journal Of Documentation*, 60(5), Pp.11–21. Available At: Http://Www.Emeraldinsight.Com/Doi/Abs/10.1108/Eb026526.

WEKA 3 – Data Mining with Open Source Machine Learning Software In Java.“ [Online]. Available: Http://Www.Cs.Waikato.Ac.Nz/Ml/Weka/.

WHIDHASIH, R.N., WAHANANI, N.A. & SUPRIYANTO, 2013. Klasifikasi Buah Belimbing berdasarkan Citra Red-Green-Blue Menggunakan KNN dan LDA. *Jurnal Penelitian Ilmu Komputer, System Embedded & Logic*, 1(1), Pp.29–35.