ANALISA PERBANDINGAN METODE NAÏVE *BAYES CLASSIFIER* DAN *K-NEAREST NEIGHBOR* TERHADAP KLASIFIKASI DATA

Aida Indriani

Teknik Informatika, STMIK PPKIA Tarakanita Rahmawati
Jl. Yos Sudarso No. 8, Tarakan, 77111
E-mail : aida@ppkia.ac.id

ABSTRAK

Penggunaan forum sebagai sarana pembelajaran telah banyak digunakan pada kalangan Mahasiswa. Forum digunakan sebagai sarana berdiskusi antar sesama anggota forum untuk membahas materi sesuai dengan judul topik. Judul topik biasanya ditentukan sesuai dengan isi materi yang akan dibahas. Judul topik yang sudah terlalu banyak di dalam sebuah forum dapat berakibat salah dalam pemilihan judul. Salah satu cara untuk mengatasinya yaitu dengan melakukan klasifikasi judul topik secara otomatis sesuai dengan isi materi. Klasifikasi teks dapat diselesaikan dengan menggunakan teknik *text mining*. Pada proses klasifikasi yang dilakukan yaitu dengan membagi *dataset* menjadi 2 bagian menjadi data latih (*training*) dan data uji (*testing*). Pada tahapan awal klasifikasi dilakukan proses *pre-processing* yang diawali dengan tahapan tokenisasi, kemudian dilanjutkan dengan *filtering* dan diakhiri dengan *stemming*. Ada beberapa metode yang dapat digunakan dalam klasifikasi teks antara lain naïve *bayes classifier* (nbc), *k-nearest neighbor* (k-nn), *rocchio*, *weight adjusted k-nearest neighbor* (wa k-nn) dan lain-lain. Pada penelitian ini, penulis membandingkan 2 metode yaitu nbc dan k-nn. Dari hasil perbandingan kedua metode dapat disimpulkan bahwa metode k-nn lebih baik tingkat akurasi nya daripada metode nbc. Hal ini dibuktikan dengan tingkat akurasi sebesar 80% untuk metode k-nn dan sebesar 73% untuk nbc yang dihitung dengan menggunakan metode *confusion matrix*.

Kata Kunci: *Klasifikasi, Forum, K-Nn, Nbc Classifier, Confusion matrix*

1. PENDAHULUAN

Forum adalah sarana yang digunakan oleh orang-orang untuk berdiskusi. Kelebihan dari forum yaitu, tidak terbatas oleh ruang dan waktu. Dimana saja, kapan saja kita membutuhkan informasi dapat diperoleh melalui forum. Mulai dari pemerintahan, hiburan, maupun pendidikan telah banyak yang menggunakan forum sebagai sarana untuk mendapatkan informasi dan berdiskusi. Dalam dunia pendidikan, telah banyak perguruan tinggi yang menggunakan forum sebagai media dalam berdiskusi antara dosen dengan mahasiswa ataupun antara mahasiswa dengan mahasiswa. Dengan adanya forum dapat membuat mahasiswa tidak perlu menunggu antrean jika ingin berdiskusi dengan dosen.
Pada sebuah forum tidak terbatas hanya membahas satu materi saja, akan tetapi dapat lebih banyak materi yang dapat dibahas. Untuk memudahkan dalam berdiskusi, forum dibedakan menjadi beberapa kelas sesuai dengan isi materi. Kelas yang banyak pada sebuah forum akan membuat si pengguna bingung dalam menentukan kelas mana yang sesuai dengan materi yang akan si pengguna diskusikan. Selain itu juga pemilihan kelas secara manual dapat membuang waktu yang digunakan dalam mencari kelas yang sesuai dengan materi.

Perbandingan 2 metode dalam klasifikasi data pernah dilakukan pada penelitian sebelumnya yang berjudul "Perbandingan Naïve *Bayes Classifier* dan *Support vector machine* untuk Klasifikasi Judul Artikel" dimana hasil dari penelitian yang diperoleh yaitu metode Naïve *Bayes Classifier* memiliki performa akurasi yang lebih baik dibanding dengan metode *Support vector machine* (Ma'arif, 2016).

Berdasarkan penelitian sebelumnya dengan judul "Klasifikasi Data Forum dengan menggunakan Metode Naïve *Bayes Classifier*" dari 15 data uji yang dilakukan diperoleh nilai akurasi dengan menggunakan *confusion matrix* sebesar 73% (Indriani, 2014). Pada penelitian ini, penulis menggunakan metode k-NN dengan pembobotan tf-idf sebagai pembanding untuk melakukan klasifikasi data forum dengan menggunakan data uji yang sama. Pembobotan tf-idf adalah pembobotan berupa angka (*numeric*) yang digunakan untuk mengetahui keterkatitan kata (*term*) terhadap dokumen (Herwijayanti dkk, 2018). Tingkat perbandingan dilihat dari nilai akurasi klasifikasi yang diperoleh dari kedua metode yang digunakan.

2. RUANG LINGKUP

Ruang lingkup yang dibahas pada penelitian ini adalah membandingkan 2 metode klasifikasi yang biasa digunakan yaitu k-NN dan NBC. Pada penelitian ini, penulis ingin mengetahui diantara kedua metode yang digunakan, metode mana yang memiliki nilai akurasi yang lebih besar. Pada penelitian sebelumnya (Indriani,

2014), penulis mendapatkan nilai akurasi sebesar 73% untuk metode NBC.

Pada penelitian ini, penulis menggunakan data forum yang berjumlah 21 data yang kemudian dibagi menjadi data latih dan data uji. Dimana data forum yang digunakan, sebelumnya telah dilakukan perhitungan pada penelitian sebelumnya (Indriani, 2014).
Dari perbandingan kedua metode dihasilkan sebuah nilai akurasi. Nilai akurasi dari kedua metode dihitung dengan menggunakan metode *Confusion matrix*. Sehingga dapat diketahui metode mana yang memiliki nilai akurasi yang lebih besar untuk kasus pengklasifikasian data forum.

3. BAHAN DAN METODE

Pada Bab ini akan menjelaskan bahan/teori yang digunakan dalam penelitian yang melingkupi formula sampai metodologi/tahapan dalam penyelesaian kasus.

3.1 *Text mining*

Klasifikasi dokumen merupakan salah satu permasalahan yang dapat diselesaikan dengan teknik *text mining*. Tujuan dari *text mining* yaitu untuk memperoleh informasi dari sekumpulan dokumen yang akan digunakan dalam proses klasifikasi itu sendiri. Selain itu, *text mining* digunakan untuk mencari kata dari sekumpulan dokumen untuk kemudian dilakukan analisis terkait hubungan kata dalam dokumen tersebut (Indriani dkk, 2018).

3.2 *Pre-processing*

Tahapan awal pada penelitian ini, yaitu melakukan *pre-processing* terhadap koleksi data forum sehingga menghasilkan kumpulan term. *Pre-processing* yang dilakukan pada penelitian ini yaitu tokenisasi, *filtering* dan *stemming*. Tokenisasi yaitu melakukan pemotongan kata dari sebuah kalimat, setelah itu dilakukan *filtering* yang digunakan untuk membuang kata yang tidak penting seperti stop list untuk kemudian dilakukan *stemming* merubah kata yang ada menjadi bentuk kata dasar (Hapsari, 2015).

3.3 *Naïve Bayes Classifier (NBC)*

Metode *naive bayes classifier* merupakan pengklasifikasian secara statistik yang dapat memprediksi probabilitas keanggotaan kelas suatu data yang masuk di dalam kelas tertentu, berdasarkan perhitungan probabilitas (Handayani dan Pribadi, 2015). Rumus untuk melakukan perhitungan perbandingan antara term data uji dengan kelas yang ada menggunakan (1).

$$P(a_i|v_j) = \frac{n_c + mp}{n + m} \quad (1)$$

Diketahui n adalah jumlah term pada data latih dimana v = v_j, n_c adalah jumlah term dimana v = v_j dan a = a_i, p adalah probabilitas setiap kelas dalam data latih dan m adalah jumlah term pada data uji. Rumus untuk menentukan kelas dari data uji menggunakan (2) (Indriani, 2014).

$$V_{nb} = argmax\ v_{i=V} P(v_i) \prod P\ (a_i|v_i) \quad (2)$$

3.4 *k-nearest Neighbor (k-NN)*

Metode pembanding yang digunakan pada penelitian ini adalah algoritma k-NN. Algoritma k-NN adalah algoritma yang melakukan klasifikasi berdasarkan k data latih yang memiliki jarak terdekat dari data uji. Jarak dihitung dengan menggunakan metode cosine similarity dengan (3) (Rivki dan Bachtiar, 2017).

$$Cos(\theta_{QD}) = \frac{\sum_{i=1}^{n} Q_i D_i}{\sqrt{\sum_{i=1}^{n} (Q_i)^2} \cdot \sqrt{\sum_{i=1}^{n} (D_i)^2}} \quad (3)$$

Diketahui Cos (Q, D) adalah kemiripan Q terhadap dokumen D, Q adalah data uji, D adalah data latih, dan n adalah banyaknya data. Pembobotan kata yang digunakan yaitu pembobotan TF-IDF Tahapan akhir pada proses klasifikasi yaitu menghitung nilai akurasi kelas. Perhitungan akurasi digunakan untuk mendapatkan persentase kecocokan kelas sebenarnya terhadap kelas yang dihasilkan secara otomatis berdasarkan model klasifikasi yang telah ada.

3.5 *Confusion matrix*

Confusion matrix adalah sebuah metode untuk melakukan perhitungan akurasi. Perhitungan berdasarkan nilai yang terdapat pada tabel 1. *confusion matrix*. Bentuk tabel *confusion matrix* ditunjukkan pada tabel 1.

Tabel 1. Tabel *Confusion matrix*

Aktual	Classified as	
	+	-
+	True positives (A)	False negatives (B)
-	False Positives (C)	True negatives (D)

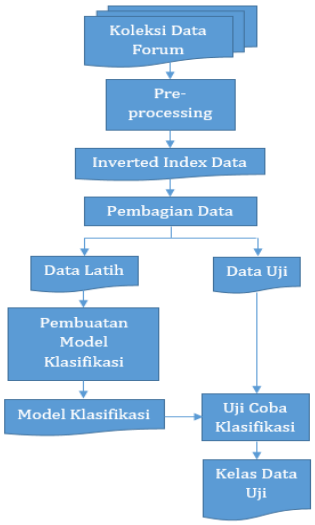
Adapun perhitungan akurasi dengan menggunakan metode *confusion matrix* seperti pada (4) (Rosandy, 2016).

$$Akurasi = \frac{A + D}{A + B + C + D}\ x\ 100\% \quad (4)$$

3.6 *Tahapan Proses Klasifikasi*

Tahapan awal yaitu melakukan *pre-processing* terhadap koleksi data forum sehingga menghasilkan kumpulan term. Dari kumpulan term yang ada, kemudian digunakan untuk membuat sebuah tabel matrix inverted index yaitu tabel kemunculan term pada masing-masing data forum. Tahapan selanjutnya yaitu melakukan

pembagian data latih dan data uji. Data latih digunakan untuk proses klasifikasi dengan menggunakan metode NBC dan k-NN sehingga mendapatkan model klasifikasi. Dari model klasifikasi yang telah diperoleh, dilakukan uji klasifikasi terhadap data uji. Tahapan akhir pada proses klasifikasi yaitu menghitung nilai akurasi kelas untuk mendapatkan persentase kecocokan kelas sebenarnya terhadap kelas yang dihasilkan secara otomatis berdasarkan model klasifikasi yang telah ada dengan menggunakan *confusion matrix*. Secara keseluruhan, metodologi penelitian ditunjukkan pada gambar 1.



Gambar 1. Tahapan Proses Klasifikasi

4. PEMBAHASAN

Hasil dari penelitian ini adalah penentuan metode klasifikasi yang lebih baik dari 2 metode yang dianalisis yaitu naïve *bayes classifier* dan *k-nearest* neighbour. Penentuan metode dengan melihat nilai akurasi dari masing-masing metode terhadap uji klasifikasi yang dilakukan.

4.1 *Pre-processing*

Pre-processing digunakan untuk mendapatkan kumpulan term dari koleksi data forum. Pada penelitian ini, data forum yang digunakan sebanyak 21 data. Beberapa data forum sesuai dengan kelas masing-masing ditunjukkan pada tabel 2.

Tabel 2. Koleksi Data Forum

No.	Data Forum	Kelas
1	Bagaimana cara membuat Galeri Image pada Eclipse	Android
2	Ada yang tau gak cara membuat koneksi pada Delphi dengan MySQL	Delphi
3	Saya kesulitan dalam membuat Mail merge pada Ms Word, bagaimana caranya ya?	Office
…	…	…
21	Bagaimana membuat form cetak dengan Delphi	Delphi

Pada tabel 2. terdapat 21 contoh data forum yang mempunyai kelas masing-masing sebanyak 7 data. Kelas dibedakan menjadi 3 kategori yaitu “Android”, “Delphi”, dan “Office”. Selanjutnya dilakukan *pre-processing* yaitu tokenisasi yang ditunjukkan pada tabel 3.

Tabel 3. Hasil Tokenisasi Data Forum

No.	Data Forum
1	bagaimana cara membuat galeri image pada eclipse
2	ada yang tau gak cara membuat koneksi pada Delphi dengan MySQL
3	saya kesulitan dalam membuat mail merge pada Ms word bagaimana caranya ya
…	…
21	bagaimana membuat form cetak dengan Delphi

Tabel 3. merupakan data forum yang telah melewati tahap pertama dalam proses *pre-processing* yaitu tokenisasi. Selanjutnya dilakukan penghapusan stop-words yang ditunjukkan pada tabel 4.

Tabel 4. Penghapusan Stop-Words Data Forum

No.	Data Forum
1	galeri image eclipse
2	koneksi Delphi MySQL
3	kesulitan mail merge Ms word
…	…
21	form cetak Delphi

Tabel 4. merupakan data forum setelah dilakukan penghapusan *stop-words*. Kata yang terdapat pada stop-*words* antara lain kata “bagaimana”, “cara”, “pada”, “saya”, “dalam” dan lain-lain. Selanjutnya yaitu tahapan *stemming* yaitu ubah term menjadi bentuk kata dasar dengan menghilangkan awalan dan akhiran dari term seperti yang ditunjukkan pada tabel 5.

Tabel 5. Hasil Stemming Data Forum

No.	Data Forum
1	galeri image eclipse
2	koneksi Delphi MySQL
3	sulit mail merge Ms word
…	…
21	form cetak Delphi

Pada tabel 5. diperoleh kumpulan kata yang telah di-*stemming*. Dari hasil *pre-processing* yang dilakukan, menghasilkan kumpulan term untuk kemudian dibuatkan sebuah tabel *matrix* yang menghubungkan antara term dengan dokumen yang dikenal dengan istilah tabel *inverted index* seperti yang ditunjukkan pada tabel 6.

Tabel 6. Tabel Inverted Index

Term	D1	D2	D3	D4	…	D21
Galeri	1	0	0	0	…	0
Image	1	0	0	0	…	0
Eclipse	1	0	0	0	…	0
Koneksi	0	1	0	0	…	0
Delphi	0	1	0	0	…	1
….	…	…	…	…	…	…
….	…	…	…	…	…	…
Android	0	0	0	0	…	0
Form	0	0	0	0	…	1
Cetak	0	0	0	0	…	1

Dari tabel 6 diperoleh jumlah masing-masing term pada setiap data yang ada. Selanjutnya dilakukan pembagian data menjadi data latih dan data uji. Pada penelitian ini menggunakan data latih sebanyak 6 yang disimbolkan sebagai "L" dan data uji sebanyak 15 yang disimbolkan sebagai "U".

4.2 *Naïve Bayes Classifier (NBC)*

Metode NBC diawali dengan tahapan yaitu penentuan nilai probabilitas pada setiap kelas terhadap data latih. Pada penjelasan sebelumnya, ditentukan jumlah data latih sebanyak 6 data yang terdiri dari 3 kelas yaitu "Android", "Office", dan "Delphi". Perhitungan probabilitas untuk kelas "Android". Probabilitas disimbolkan sebagai p.

(5)

$$p(Android) = \frac{jumlah\ kelas\ android}{jumlah\ data\ latih} = \frac{2}{6} = 0{,}33$$

Dari perhitungan probabilitas untuk kelas "Android", diperoleh nilai probabilitas yaitu 0,33. Untuk nilai probabilitas kelas "Office" dan "Delphi" dilakukan perhitungan yang sama seperti pada perhitungan probabilitas kelas "Android". Dari perhitungan yang dilakukan, diperoleh nilai probabilitas untuk kelas "Office" sebesar 0,33 dan kelas "Delphi" sebesar 0,33.

Langkah selanjutnya yaitu melakukan pengklasifikasian data uji dengan menggunakan nilai probabilitas yang telah diperoleh pada setiap kelas. Data uji sebanyak 15 dapat dilihat pada tabel 7.

Tabel 7. Contoh Koleksi Data Uji Forum

No.	Data Forum	Kelas
U1	Cara memasukkan gambar pada excel gimana ya	?
U2	Bagaimana penggunaan timer pada aplikasi Delphi	?
U4	Bagaimana membuat daftar isi secara otomatis menggunakan Ms Word	?
…	….	…
U15	Cara cepat untuk melakukan edit gambar pada Ms Power point	?

Seperti halnya data latih, data uji juga melalui tahapan *pre-processing* pada tahapan awal klasifikasi. Data uji yang telah melewati tahapan pre-processing, ditunjukkan pada tabel 8.

Tabel 8. *Pre-processing* Data Uji Forum

No.	Data Forum	Kelas
U1	masuk gambar excel	?
U2	timer aplikasi Delphi	?
U4	daftar isi otomatis Ms word	?
…	…	…
U15	edit gambar Ms PowerPoint	?

Tabel 8. merupakan data uji yang telah melalui tahapan *pre-processing* untuk kemudian akan diklasifikasikan dengan menggunakan NBC. Contoh nilai-nilai untuk kelas "Android".

Android		
Term "daftar"	Term "isi"	Term "otomatis"
n = 6	n = 6	n = 6
n_c = 0	n_c = 0	n_c = 0
p = 0,33	p = 0,33	p = 0,33
m = 5	m = 5	m = 5
Term "Ms"	Term "word"	
n = 6	n = 6	
n_c = 0	n_c = 0	
p = 0,33	p = 0,33	
m = 5	m = 5	

Dari nilai-nilai yang telah diperoleh dan dengan menggunakan (6).

(6)

$$P(Android|Daftar) = \frac{0 + 5.0{,}33}{6+5} = \frac{1{,}65}{11} = 0{,}15$$

$$P(Android|Isi) = \frac{0 + 5.0{,}33}{6+5} = \frac{1{,}65}{11} = 0{,}15$$

$$P(Android|Otomatis) = \frac{0 + 5.0{,}33}{6+5} = \frac{1{,}65}{11} = 0{,}15$$

$$P(Android|Ms) = \frac{0 + 5.0{,}33}{6+5} = \frac{1{,}65}{11} = 0{,}15$$

$$P(Android|Word) = \frac{0 + 5.0{,}33}{6+5} = \frac{1{,}65}{11} = 0{,}15$$

Untuk perhitungan $P(a_i|v_j)$ yang lainnya dilakukan proses yang sama seperti pada perhitungan P(Android|Daftar). Dengan menggunakan (7), yaitu mencari nilai maksimal dari hasil perkalian nilai probabilitas dan nilai P untuk setiap kelas, diperoleh hasil.

(7)

V(Android)=0,33*0,15*0,15*0,15*0,15*0,15 = 0,000026
V(Delphi)=0,33*0,15*0,15*0,15*0,15*0,15 = 0,000026
V(Office)=0,33*0,13*0,13*0,13*0,28*0,21=0,000040
Vnb = argmax (v(android) | v(Delphi) | v(office))
Vnb = argmax (0,000026 | 0,000026 | 0,000040)

Vnb = 0,000040

Nilai maksimal yang diperoleh yaitu 0,000040. 0,000040 nilai v untuk kelas "Office". Jadi kesimpulan yang diperoleh adalah data uji U4 termasuk kelas "Office". Untuk data uji yang lain dilakukan proses yang sama seperti pada data uji U4.

4.3 *k-nearest Neighbour*

Proses klasifikasi k-NN diawali dengan melakukan pembobotan masing-masing term terhadap data latih dan data uji. Pembobotan yang dilakukan dalam klasifikasi dengan menggunakan tf-idf. Sebagai contoh, data uji ke-1 yang telah mengalami proses *pre-processing* yaitu "daftar isi otomatis ms word". Langkah awal dari pembobotan term yaitu menghitung nilai idf dengan menggunakan rumus log(D/df) +1. Hasil perhitungan idf untuk masing-masing term ditunjukkan pada tabel 9.

Tabel 9. Hasil Perhitungan IDF

Term	L1	L2	L3	L4	L5	U1	df	idf
Galeri	1	0	0	0	0	0	1	1.845
Image	1	0	0	0	0	0	1	1.845
Eclipse	1	0	0	0	0	0	1	1.845
Koneksi	0	1	0	0	0	0	1	1.845
Delphi	0	1	0	0	0	0	2	1.544
….	…	…	…	…	…	…	…	…
android	0	0	0	0	1	0	1	1.845
ms	0	0	1	1	0	1	3	1.368
word	0	0	1	0	0	1	2	1.544
form	0	0	0	0	0	0	1	1.845
cetak	0	0	0	0	0	0	1	1.845

Pada tabel 9, untuk term "galeri" menghasilkan nilai df = 1 yang diperoleh dari jumlah dokumen yang mengandung term "galeri". D adalah jumlah keseluruhan dokumen, D = 7. Untuk hasil idf = log (7/1) + 1 = 1.845. setelah mendapatkan nilai idf, selanjutnya yaitu menghitung bobot setiap term dengan rumus tf x idf. Hasil perhitungan tf-idf ditunjukkan pada tabel 10.

Tabel 10. Hasil Perhitungan TF-IDF

Term	L1	L2	L3	L4	U1
galeri	1.845	0	0	0	0
image	1.845	0	0	0	0
eclipse	1.845	0	0	0	0
koneksi	0	1.845	0	0	0
Delphi	0	1.544	0	0	0
….	…	…	…	…	…
android	0	0	0	0	0
Ms	0	0	1.368	1.368	1.368
word	0	0	1.544	0	1.544
form	0	0	0	0	0
cetak	0	0	0	0	0

Pada tabel 10, untuk term "galeri" pada dokumen "L1" diperoleh nilai bobot sebesar 1.845 yang didapat dari nilai tf yaitu jumlah term "galeri" pada dokumen "L1" sebanyak 1 kemudian dikalikan dengan nilai idf term "galeri" sebesar 1.845 seperti yang ditunjukkan pada tabel 9. Setelah bobot masing-masing term diperoleh, selanjutnya dilakukan perhitungan cosine similarity untuk mendapatkan jarak antara dokumen data latih dan data. Nilai total bobot dokumen data latih ditunjukkan pada tabel 11.

Tabel 11. Nilai Total Bobot Data Latih

Term	L1	L2	L3	L4	L5	L6
galeri	0	0	0	0	0	0
image	0	0	0	0	0	0
eclipse	0	0	0	0	0	0
koneksi	0	0	0	0	0	0
Delphi	0	0	0	0	0	0
….	…	…	…	…	…	…
android	0	0	0	0	0	0
Ms	0	0	1.871	1.871	0	0
word	0	0	2.384	0	0	0
form	0	0	0	0	0	0
cetak	0	0	0	0	0	0
Total	0	0	4.255	1.871	0	0

Pada tabel 11, untuk dokumen data latih "L3" mendapatkan nilai bobot sebesar 4.255. Langkah selanjutnya yang dilakukan yaitu melakukan perhitungan untuk mendapatkan akar dari total nilai bobot setiap dokumen data latih dan data uji dengan terlebih dahulu melakukan perhitungan nilai bobot setiap term yang dipangkatkan. Hasil perhitungan untuk setiap dokumen data latih dan data uji ditunjukkan pada tabel 12.

Tabel 12. Hasil Perhitungan Akar Nilai Bobot

Term	L1	L2	L3	L4	U1
galeri	3.404	0	0	0	0
image	3.404	0	0	0	0
eclipse	3.404	0	0	0	0
koneksi	0	3.404	0	0	0
Delphi	0	2.384	0	0	0
….	…	…	…	…	…
android	0	0	0	0	0
Ms	0	0	1.871	1.871	1.871
word	0	0	2.384	0	2.384
form	0	0	0	0	0
cetak	0	0	0	0	0
Total	10.213	9.193	11.064	12.085	14.469
SQRT	3.196	3.032	3.326	3.476	3.804

Pada tabel 12, untuk dokumen data latih "L1" diperoleh nilai bobot setelah di akarkan sebesar 3.196 dan data uji "U1" sebesar 3.804. Langkah selanjutnya yaitu mendapatkan nilai jarak antara dokumen data latih dan data uji. Nilai bobot "L3" pada tabel 9 dengan nilai 4.255, nilai akar bobot "L3" pada tabel 10 dengan nilai 3.326 dan nilai akar bobot "U1" pada tabel 10 dengan nilai 3.804. dari nilai-nilai tersebut dilakukan perhitungan yaitu 4.255/(3.326*3.804) = 0.336. Nilai jarak untuk setiap dokumen data latih terhadap data uji ditunjukkan pada tabel 13.

Tabel 13. Hasil Perhitungan Jarak

Dokumen	$Cos(\theta_{..})$	Jarak Terdekat	Kelas Forum
L1	0	3	Android
L2	0	4	Delphi
L3	0.336	1	Office
L4	0.142	2	Office
L5	0	5	Android
L6	0	6	Delphi

4.4 *Confusion matrix*

Confusion matrix digunakan untuk pengukuran efektivitas klasifikasi. Dengan menggunakan 4, dilakukan proses perhitungan akurasi untuk 15 data uji. Hasil klasifikasi dengan k-NN untuk 15 data uji ditunjukkan pada tabel 14.

Tabel 14. Hasil Kelas Data Uji Metode k-NN

No.	Kelas Sebenarnya	Kelas k-NN
U1	Office	Office
U2	Android	Android
U3	Office	Android
U4	Delphi	Delphi
U5	Delphi	Delphi
U6	Office	Office
U7	Android	Android
U8	Android	Android
U9	Delphi	Delphi
U10	Office	Delphi
U11	Office	Office
U12	Android	Android
U13	Delphi	Delphi
U14	Android	Delphi
U15	Delphi	Delphi

Dari tabel 14, diperoleh nilai A = 3, D = 9, B = 2, C = 1. Langkah berikutnya yaitu dengan menggunakan (8), dihitung nilai akurasi.

(8)

$$Akurasi = \frac{3+9}{3+9+2+1} \; x\; 100\% = \frac{12}{15} = 80\%$$

Kesimpulan yang diperoleh yaitu, dari 15 data uji diperoleh akurasi kecocokan kelas sebenarnya terhadap kelas prediksi dengan k-NN sebesar 80%.

4.5 Hasil Analis Perbandingan

Berdasarkan penelitian sebelumnya dengan judul "Klasifikasi Data Forum dengan menggunakan Metode Naïve *Bayes Classifier*" dari 15 data uji yang dilakukan diperoleh hasil uji klasifikasi seperti pada tabel 15.

Tabel 15. Hasil Kelas Data Uji Metode NBC

No.	Kelas Sebenarnya	Kelas NBC
U1	Office	Office
U2	Android	Android
U3	Office	Android
U4	Delphi	Delphi
U5	Delphi	Delphi
U6	Office	Office
U7	Android	Office
U8	Android	Android
U9	Delphi	Delphi
U10	Office	Delphi
U11	Office	Office
U12	Android	Android
U13	Delphi	Delphi
U14	Android	Delphi
U15	Delphi	Delphi

Dari tabel 15, diperoleh nilai A = 3, D = 8, B = 2, C = 2. Langkah berikutnya yaitu dengan menggunakan (9), dihitung nilai akurasi .

$$Akurasi = \frac{3+8}{3+8+2+2} \; x100\% = \frac{11}{15} = 73\% \quad (9)$$

Kesimpulan yang diperoleh yaitu, dari 15 data uji diperoleh akurasi kecocokan kelas sebenarnya terhadap kelas prediksi dengan NBC sebesar 73%. Hasil analisis dari perbandingan kedua metode yaitu metode k-NN memiliki akurasi yang lebih baik dibanding dengan metode NBC dengan nilai akurasi sebesar 80% untuk metode k-NN dan 73% untuk metode NBC.

5. KESIMPULAN

Penggunaan metode NBC *dank*-NN dapat digunakan untuk pengklasifikasian otomatis terhadap data forum, dimana metode NBC dengan menggunakan probabilitas kemunculan kata, sedangkan untuk k-NN dengan melihat kedekatan jarak antara dokumen data latih terhadap data uji. Pengukuran efektivitas klasifikasi terhadap 15 data uji diperoleh nilai 80% untuk metode k-NN dan sebesar 73% untuk metode NBC dengan menggunakan pengukuran efektivitas *Confusion matrix*. Berdasarkan nilai efektivitas klasifikasi, dapat disimpulkan bahwa metode k-NN lebih baik dibanding metode NBC dikarenakan k-NN memiliki nilai efektivitas lebih besar dibanding NBC.

6. SARAN

Untuk perkembangan lebih lanjut dapat melakukan perbandingan dengan metode lainnya seperti WA k-NN yang merupakan pengembangan dari metode k-NN.

7. DAFTAR PUSTAKA

Handayani & Pribadi, F.S. 2015. Implementasi Algoritma Naïve *Bayes Classifier* dalam Pengklasifikasian Teks Otomatis Pengaduan dan Pelaporan Masyarakat melalui Layanan Call Center 110. Jurnal Teknik Elektro. Vol. 7 No. 1, pp. 19-24.

Hapsari, I.T., Andoko, B.S., & Rahmad, C. 2015. Aplikasi Information Retrieval untuk Pencarian Dokumen Laporan Penelitian. Jurnal Informatika Polinema. ISSN: 2407-070x Vol. 1 No. 3, pp. 23-28.

Herwijayanti, B., Ratnawati, D.E., & Muflikhah, L. 2018. Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity. Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. e-ISSN: 2548-964X Vol. 2, No.1 hlm:306-312.

Indriani, A. 2014. Klasifikasi Data Forum dengan menggunakan Metode Naïve *Bayes Classifier*. Seminar Nasional Aplikasi Teknologi Informasi (SNATI). ISSN: 1907-5022, pp. G5-G10.

Indriani, A., Muhammad, M., Suprianto, S., & Hadriansa, H. 2018. Implementasi Jaccard Index dan N-gram pada Rekayasa Aplikasi Koreksi Kata Berbahasa Indonesia. Jurnal Sebatik. Vol. 22 No. 2, pp. 95-101.

Ma'arif, M.R. 2016. Perbandingan Naïve *Bayes Classifier* dan *Support vector machine* untuk Klasifikasi Judul Artikel. Jurnal Informatika Sunan Kalijaga. ISSN: 2527-5836 Vol. 1 No. 2 pp. 90-93.

Rivki, M., & Bachtiar, A.M. 2017. Implementasi Algoritma *k-nearest Neighbor* dalam Pengklasifikasian Follower Twitter yang menggunakan Bahasa Indonesia. Jurnal Sistem Informasi. Vol. 13 Issue 1, pp. 31-37.

Rosandy, T. 2016. Perbandingan Metode Naïve *Bayes Classifier* dengan Metode Decision Tree (c4.5) untuk menganalisa Kelancaran Pembiayaan (study kasus: kspps/bmt al-fadhila). Jurnal TIM Darmajaya. ISSN: 2442-5567 Vol. 2 No. 1, pp. 52-62