Jurnal Infomedia Vol. 4 No. 2 Desember 2019 | P-ISSN: 2527-9858
E-ISSN: 2548-1180

IMPLEMENTASI METODA *NAÏVE BAYES* DAN *VECTOR SPACE MODEL* DALAM DETEKSI KESAMAAN ARTIKEL JURNAL BERBAHASA INDONESIA

Ni Luh Wiwik Sri Rahayu Ginantra[1], Ni Wayan Wardani[2]

[1,2] Jurusan Teknik Informatika STMIK STIKOM Indonesia
[1] wiwik@stiki-indonesia.ac.id , [2] niwayan.wardani@stiki-indonesia.ac.id

Abstrak— Salah satu cara untuk menjaga kualitas karya ilmiah di Indonesia adalah dengan memeriksa artikel sebelum dipublikasikan. Pengecekan sebelum publikasi dilakukan agar tingkat kesamaan tidak tinggi karena makalah yang diterbitkan dapat dikutip untuk menyebabkan tingkat kesamaan yang tinggi. Masalah selanjutnya adalah pentingnya pengelompokkan makalah topik, dimana makalah yang akan diperiksa harus memiliki kategori yang sama dengan makalah yang dibandingkan. Data penelitian dalam mengkategorikan jurnal dan pengujian diambil dari situs *neliti.com*. Dalam penelitian ini, untuk mengklasifikasikan topik jurnal menggunakan algoritma *Naive Bayes* dan untuk mengukur kesamaan makalah menggunakan metoda *Vector Space Model*. Algoritma *Naive Bayes* tidak dapat mengklasifikasikan ke dalam satu topik jurnal dengan tepat tetapi mengklasifikasikan menjadi beberapa topik jurnal sehingga mempengaruhi kinerja metoda *Vector Space Model*. Hasil perhitungan deteksi kesamaan teks oleh *Vector Space Model* dapat mencapai 90% ke atas untuk data uji tertentu. Hasil perhitungan deteksi kesamaan teks dengan metoda *Vector Space Model* juga sangat dipengaruhi oleh data pelatihan. Semakin lengkap dan kompleks data pelatihan, maka semakin valid hasil pengujian kinerja *Vector Space Model*.
Kata kunci— Text Similarity, Naive Bayes, VSM.

Abstract— One way to maintain the quality of scientific work in Indonesia is by checking articles before they are published. Checking before the publication was done so that the similarity level is not high because the published papers can be quoted to cause a high level of similarity. The next problem is the importance of grouping topic papers, where papers to be checked should have the same category as comparative papers. Research data in categorizing journals and testing is taken from the neliti.com site. In this study, to classify the topic of the journal using the Naïve Bayes algorithm and to measure the similarity of papers using the Vector Space Model method. Naïve Bayes algorithm cannot classify into one journal topic precisely but classifies it into several journal topics so that it affects the performance of the Vector Space Model method. The results of the calculation of text similarity detection by the Vector Space Model can reach 90% and above for test data. The results of the calculation of text similarity detection by the Vector Space Model are also strongly influenced by training data. The more complete and complex of the training data, then more valid the results of the Vector Space Model performance testing.
Keywords— Text similarity, Naïve Bayes, VSM.

I. PENDAHULUAN

Saat ini, salah satu poin penting dalam menjalankan fungsi Tridharma Perguruan Tinggi oleh dosen adalah melaksanakan penelitian dan mempublikasikan hasil pemikiran serta analisisnya tersebut. Kinerja dosen yang selanjutnya menjadi kinerja jurusan, fakultas dan perguruan tinggi sangat dipengaruhi oleh seberapa luas dan berkualitasnya publikasi para dosen tetapnya.

Tuntutan publikasi yang dilakukan komunitas akademik perguruan tinggi memberikan dampak yang cukup besar terhadap kesadaran para dosen pentingnya melakukan kajian, penelitian serta menulis karya ilmiah. Perkembangan karya ilmiah di Indonesia relative makin baik, terutama sejak diberlakukannya regulasi pemerintah, yang mewajibkan mahasiswa S1, S2 hingga S3 untuk menulis artikel di jurnal ilmiah sebagai salah satu prasyarat kelulusan. Bagi dosen tentunya akan semakin besar tuntutannya untuk aktif menulis di jurnal ilmiah baik di tingkat nasional terakreditasi maupun jurnal internasional bereputasi.

Sejalan dengan regulasi pemerintah tersebut, maka akan terjadi peningkatan kuantitas publikasi karya ilmiah oleh kalangan akademisi. Dengan semakin meningkatnya jumlah publikasi maka kualitas karya ilmiah juga sangat penting diperhatikan. Salah satu cara menjaga kualitas karya ilmiah akademisi di Indonesia adalah dengan mengecek artikel sebelum dipublikasikan. Pengecekan di awal sebelum terbit dilakukan agar tingkat *similarity* tidak tinggi, karena makalah yang sudah terbit dapat dikutip sehingga menyebabkan kadar *similarity* tinggi.

Selain perlunya mengecek artikel sebelum dipublikasikan, permasalahan berikutnya adalah pentingnya pengelompokkan topik makalah, dimana makalah yang akan di cek sebaiknya memiliki kategori yang sama dengan artikel – artikel pembanding.

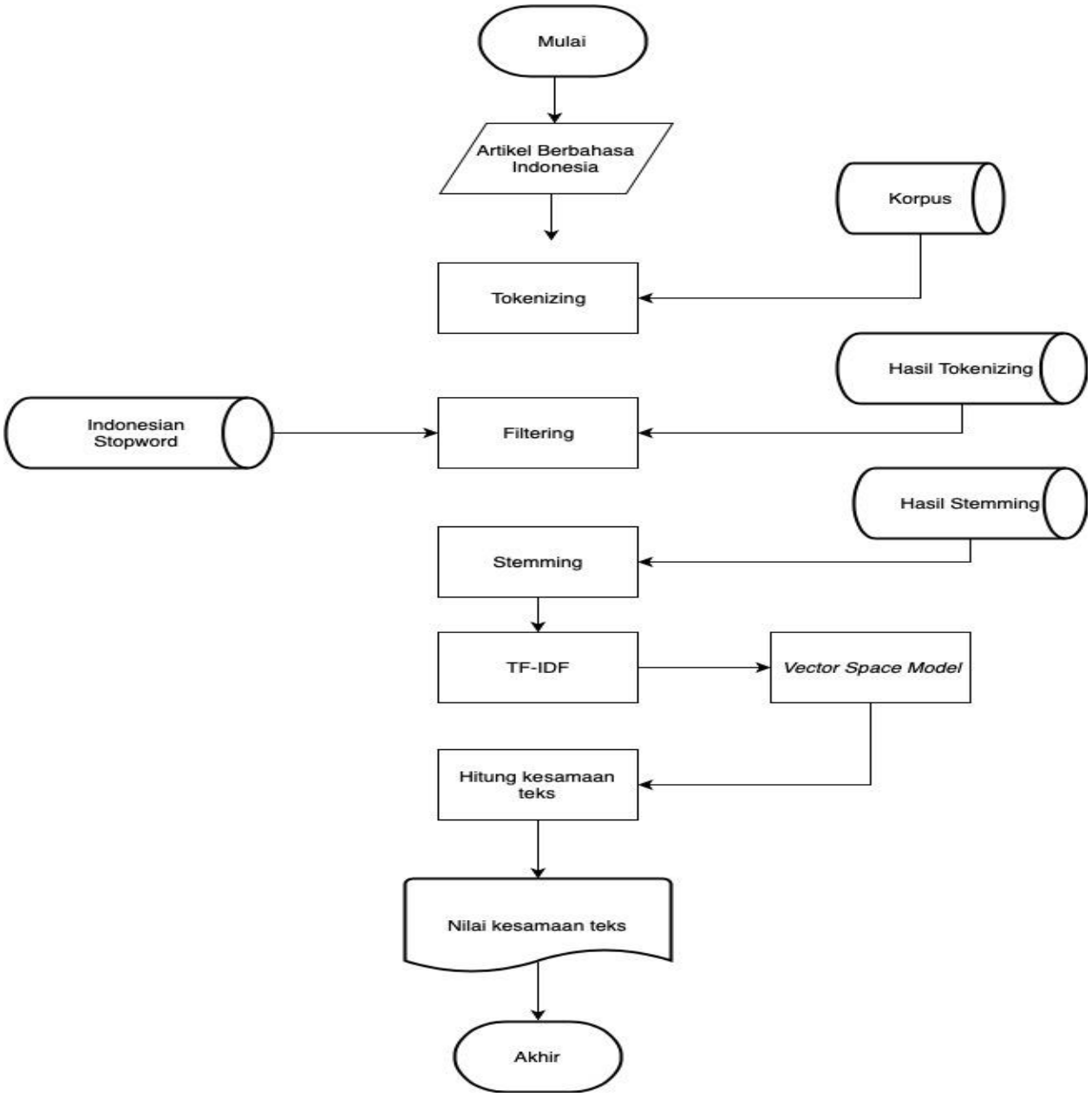
Berdasarkan masalah tersebut, maka pada penelitian ini akan digunakan algoritma *Naïve Bayes* untuk klasifikasi artikel – artikel ke dalam satu topik. Cara kerja dari algoritma *Naïve Bayes* adalah menggunakan perhitungan probabilitas. Konsep dasarnya adalah menghitung peluang dari suatu kelas dari masing – masing kelompok atribut yang ada, dan menentukan kelas mana yang paling optimal. Proses pengelompokkan atau klasifikasi dibagi ke dalam dua fase yaitu *learning* / *training* dan *testing* / *classify*. Pada fase *learning* sebagian data yang telah diketahui kelas datanya, diumpamakan untuk membentuk model perkiraan, kemudia pada fase *testing model* yang sudah terbentuk diuji dengan sebagian data.

Selanjutnya untuk mengecek kesamaan teks pada makalah akan menggunakan metoda *Vector Space Model* (VSM) yang merupakan sebuah pendekatan natural dari setiap kata dalam suatu dimensi spasial. Dokumen dipandang sebagai sebuah vektor yang memiliki *magnitude* (jarak) dan *direction* (arah). Pada VSM, sebuah kata direpresentasikan dengan sebuah dimensi dari ruang vektor. Relevansi sebuah dokumen ke

sebuah kueri didasarkan pada similaritas diantara vektor dokumen dan vektor kueri..

II. METODOLOGI PENELITIAN

Berikut adalah diagram dari rancangan penelitian :



Gambar 1. Rancangan Penelitian.

1. Dataset yang digunakan dalam penelitian ini adalah dokumen PDF dalam Bahasa Indonesia yang berasal dari repositori jurnal neliti.com. Topik jurnal yang digunakan dalam penelitian ini juga mengambil topik jurnal yang terdapat di neliti.com sebanyak 67 topik atau bidang studi [1].

2. *Text Processing*

Text Mining adalah proses pengambilan data dalam bentuk teks dari sumber; dalam hal ini, sumbernya adalah dokumen. Dengan *text mining*, dapat mencari kata kunci yang dapat mewakili konten suatu dokumen dan kemudian menganalisis dan melakukan pencocokan antara dokumen dan kata kunci basis data yang telah dibuat. *Text Processing* adalah bagian dari *text mining*. Tahap – tahap *text processing* secara umum adalah tokenisasi, *stopword*, dan *stemming* [2].

a. *Tokenisasi*

Tokenisasi adalah proses yang dilakukan pada dokumen untuk mendapatkan persyaratan. Proses yang dilakukan adalah memotong kata – kata yang membangun dokumen, dan hasil potongan disebut token, dan mungkin dalam proses yang sama melemparkan berbagai karakter[3].

b. *Stopword*

Stopword adalah proses yang dilakukan setelah *tokenizing* pada pemrosesan teks. Proses berhenti adalah menghilangkan kata – kata yang sering muncul secara umum, disebut *stopwords*. *Stopword* cenderung memiliki bobot yang rendah, sehingga hampir tidak mempengaruhi perhitungan jika *stopword* dihapus. Salah satu Teknik yang biasa digunakan untuk mengurangi indeks kata adalah dengan membendung atau menghapus kata kunci[4].

c. Stemming

Stemming adalah proses mendapatkan kata dasar dari suatu istilah. Tujuan dari proses ini dilakukan agar makna suatu istilah dari satu dokumen sama dengan dokumen lainnya karena istilah tersebut sudah dalam bentuk dasar. Untuk alasan transformasi kata, dokumen biasanya menggunakan bentuk kata yang berbeda, meskipun kata tersebut memiliki makna yang tidak jauh berbeda. Dalam banyak situasi, akan sangat membantu jika bentuk kata yang berbeda dianggap sama.

3. *Term Frequency-Inversed Document Frequency Algorithm* (TF-IDF)

Algoritma TF-IDF adalah suatu algoritma yang berdasarkan nilai statistik menunjukkan kemunculan suatu kata di dalam dokumen. Dalam penelitian (Nengsih, 2014) menurut Feldman *TF* (*Term Frequency*) menyatakan banyaknya suatu kata muncul dalam sebuah dokumen dan *DF* (*Document Frequency*) menyatakan banyaknya dokumen yang mengandung suatu kata dalam satu segmen publikasi. TF-IDF adalah nilai bobot dari suatu kala yang diambil dari nilai *IF* dan nilai Inverse *DF*, yang didefinisikan [5]:

$$IDF(w) = log\left(\frac{N}{DF(w)}\right) \quad (1)$$

$$TF - IDF(w,d) = TF(w,d) \times IDF(w) \quad (2)$$

Keterangan :

TF-IDF(*w*,*d*)	:	Bobot suatu kata dalam keseluruhan dokumen
w	:	Suatu kata (word)
d	:	Suatu dokumen (document)
TF(*w*,*d*)	:	Frekuensi kemunculan sebuah kata w dalam dokumen d
IDF(*w*)	:	Inverse DF dari kata w
N	:	Jumlah keseluruhan dokumen
DF(*w*)	:	Jumlah dokumen yang menggabungkan kata w

4. *Naïve Bayes Classifier*

Naïve Bayes merupakan sebuah pengklasifikasian probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Algoritma menggunakan teorema *Bayes* dan mengasumsikan semua atribut independen atau tidak saling ketergantungan yang diberikan oleh nilai pada variable kelas. Definisi lain mengatakan *Naïve Bayes* merupakan pengklasifikasian dengan metoda probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya.

Naïve Bayes didasarkan pada asumsi penyederhanaan bahwa nilai atribut secara kondisional saling bebas jika diberikan nilai output. Dengan kata lain, diberikan nilai output, probabilitas mengamati secara bersama adalah produk dari probabilitas individu. Keuntungan penggunaan *Naïve Bayes* adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (*Training Data*) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. *Naïve Bayes* sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks daripada yang diharapkan, maka metoda *Naïve Bayes* adalah metoda yang dipergunakan untuk proses klasifikasi teks dalam penelitian ini. Terdapat 2 tahap pada proses klasifikasi teks. Tahap pertama adalah pelatihan terhadap himpunan artikel contoh (*training example*), sedangkan tahap kedua adalah proses klasifikasi dokumen yang belum diketahui topiknya.
Teorema Bayes :

$$P(Ci|X) = \frac{P(X|Ci) \times P(Ci)}{P(X)} \quad (3)$$

Keterangan :

P(Ci|X)	:	Probabilitas kemunculan kelas *Ci* dengan *X P(X)* "konstan" untuk semua kelas sehingga hanya terbentuk *P(X|Ci) × P(Ci)* yang perlu dimaksimumkan
X	:	Kejadian *X*
Ci	:	Kelas yang tersedia *(C1, C2, …Ci)*
P(Ci)	:	Probabilitas kemunculan kelas *Ci*
P(X)	:	Probabilitas kemunculan kejadian *X*
P(X|Ci)	:	Probabilitas kemunculan kejadian *X* dengan kondisi *Ci*

$$P(X|Ci) = P(Xt|Ci) \quad (4)$$

Keterangan :

Xt	:	Nilai – nilai atribut dalam sample *X*
P(Xt|Ci)	:	Probabilitas kejadian *Xt* dengan kondisi *Ci*

dapat dihitung dari database training

5. *Vector Space Model*

Metoda *Vector Space Model* atau *Term Vector Model* adalah sebuah model aljabar untuk menggambarkan dokumen teks (beberapa objek) sebagai *vector* dari *identifier*. Biasanya digunakan dalam penyaringan informasi (*information retrieval*), *indexing* dan pemberian ranking yang saling relevan. Proses dari perhitungan metoda ini adalah *indexing* dokumen, pembobotan *term* dan perhitungan kesamaan. Proses *indexing* dokumen adalah proses melalui tahapan – tahapan dalam *text mining*. Proses selanjutnya adalah pembobotan term dengan menggunakan algoritma TF/DF. Proses yang terakhir adalah perhitungan kesamaan dengan pendekatan *Cosine*, yang dinyatakan dalam rumus [6]:

$$Similarity(dj, qk) = \frac{\sum_{i=1}^{n}(tdjj \times tqik)}{\sqrt{\sum_{i=1}^{n} tdjj \times \sum_{i=1}^{n} tqik}} \quad (10)$$

Keterangan :

Similarity(dj,qk) : Tingkat kesamaan suatu dokumen dengan query tertentu
tdij : Term ke-i dalam vector untuk dokumen ke-*j*
tqik : Term ke-i dalam vector untuk query ke-*k*
n : Jumlah term yang unik dalam dataset

6. *Klasifikasi Topik Jurnal dengan Naïve Bayes*

Tahap klasifikasi topik jurnal dengan *Naïve Bayes* adalah sebagai berikut :

TABEL I
CONTOH DOKUMEN

1	Penelitian ini berupa pengembangan Sistem Pendukung Keputusan (SPK) untuk perencanaan kebutuhan bahan baku yang mempunyai batas masa kadaluarsa dan adanya ketentuan diskon bagi pembelian dalam jumlah tertentu
2	Pemeriksaan pajak merupakan serangkaian kegiatan untuk mencari, mengumpulkan, mengolah data dana tau keterangan lainnya untuk menguji kepatuhan pemenuhan kewajiban perpajakan dan untuk tujuan lain dalam rangka melaksanakan ketentuan

1. S*topword Removal*

Penghapusan konjungsi yang ada dalam dokumen dan menghitung frekuensi terjadinya konjungsi yang akan dihapus dalam contoh dokumen.

TABEL 2
STOPLIST

No	Stoplist	Frekuensi
1	ini	1
2	berupa	1
3	untuk	3
4	yang	1
5	dan	3
6	adanya	1
7	bagi	1
8	dalam	2
9	tertentu	1
10	mencari	1
11	atau	1
12	lainnya	1
13	lain	1

2. *Tokenisasi*

Setelah melakukan penghapusan kata penghubung. Contoh dokumen yang sudah dihapus kata penghubungnya dan diubah semua huruf besarnya setelah kumpulan karakter dalam suatu dokumen ke dalam satuan kata.

TABEL 3
DOKUMEN SETELAH DILAKUKAN STOPWORD REMOVAL

1	Penelitian berupa pengembangan Sistem Pendukung Keputusan SPK perencanaan kebutuhan bahan baku mempunyai batas masa kadaluarsa ketentuan diskon pembelian jumlah tertentu
2	Pemeriksaan pajak merupakan serangkaian kegiatan mencari mengumpulkan mengolah data keterangan menguji kepatuhan pemenuhan kewajiban perpajakan tujuan rangka melaksanakan ketentuan peraturan perundang undangan perpajakan.

TABEL 4
MENGUBAH SEMUA HURUF BESAR MENJADI KECIL

1	penelitian berupa pengembangan sistem pendukung keputusan spk perencanaan kebutuhan bahan baku mempunyai batas masa kadaluarsa ketentuan diskon pembelian jumlah tertentu
2	pemeriksaan pajak merupakan serangkaian kegiatan mencari mengumpulkan mengolah data keterangan menguji kepatuhan pemenuhan kewajiban perpajakan tujuan rangka melaksanakan ketentuan peraturan perundang undangan perpajakan

TABEL 5
PROSES TOKENISASI

No	Kata
1	penelitian
2	berupa
3	pengembangan
4	sistem
5	pendukung
6	keputusan
7	spk
8	perencanaan
9	kebutuhan
10	bahan
11	baku
12	mempunyai
13	batas
14	masa
15	kadaluarsa
16	ketentuan
17	diskon
18	pembelian
19	jumlah
20	tertentu
21	pemeriksaan
22	pajak
23	merupakan
24	serangkaian
25	kegiatan
26	mencari
27	mengumpulkan
28	mengolah
29	data
30	keterangan
31	menguji
32	kepatuhan
33	pemenuhan
34	kewajiban
35	perpajakan
36	tujuan
37	rangka
38	melaksanakan
39	ketentuan
40	peraturan
41	perundang
42	undangan
43	perpajakan

3. *Menentukan Nilai IDF*

Setelah dilakukan tokenisasi, hasil dari tokenisasi dilakukan pengecekan data pada setiap topik untuk melihat kemunculan kata, kemudian kemunculan kata tersebut (*df*) dijadikan acuan dalam mencari nilai dari *idf* dengan rumus *log*(banyaknya topik / df pada setiap kata).

TABEL 6
MENENTUKAN NILAI IDF

Kemunculan Kata	d 1	d 2	d 3	d 4	d 5	df	idf
penelitian	1	0	0	0	0	1	0,69897
berupa	1	0	0	0	0	1	0,69897
pengembangan	1	0	0	0	0	1	0,69897
Sistem	1	0	0	0	0	1	0,69897
pendukung	1	0	0	0	0	1	0,69897
keputusan	1	0	0	0	0	1	0,69897
spk	1	0	0	0	0	1	0,69897
perencanaan	1	0	0	0	0	1	0,69897
kebutuhan	1	0	0	0	0	1	0,69897
Bahan	1	0	0	0	0	1	0,69897
Baku	1	0	0	0	0	1	0,69897
mempunyai	1	0	0	0	0	1	0,69897
Batas	1	0	0	0	0	1	0,69897
Masa	1	0	0	0	0	1	0,69897
kadaluarsa	1	0	0	0	0	1	0,69897
ketentuan	1	0	0	0	0	1	0,69897
Diskon	1	0	0	0	0	1	0,69897
pembelian	1	0	0	0	0	1	0,69897
Jumlah	1	0	0	0	0	1	0,69897
Tertentu	0	0	1	0	0	1	0,69897
pemeriksaan	0	0	1	0	0	1	0,69897
Pajak	0	0	1	0	0	1	0,69897
merupakan	0	0	1	0	0	1	0,69897
serangkaian	0	0	1	0	0	1	0,69897
kegiatan	0	0	1	0	0	1	0,69897
mencari	0	0	1	0	0	1	0,69897
mengumpulkan	0	0	1	0	0	1	0,69897
mengolah	0	0	1	0	0	1	0,69897
data	0	0	1	0	0	1	0,69897
keterangan	0	0	1	0	0	1	0,69897
menguji	0	0	1	0	0	1	0,69897
kepatuhan	0	0	1	0	0	1	0,69897
pemenuhan	0	0	1	0	0	1	0,69897
kewajiban	0	0	1	0	0	1	0,69897
perpajakan	0	0	1	0	0	1	0,69897
tujuan	0	0	1	0	0	1	0,69897
rangka	0	0	1	0	0	1	0,69897
melaksanakan	0	0	1	0	0	1	0,69897
ketentuan	0	0	1	0	0	1	0,69897
peraturan	0	0	1	0	0	1	0,69897
perundang	0	0	1	0	0	1	0,69897
undangan	0	0	1	0	0	1	0,69897
perpajakan	0	0	1	0	0	1	0,69897

4. *Menentukan Nilai TD-IDF*

Hasil dari idf kemudian dicarikan nilai tf-idf dengan rumus (kemunculan kata tiap topik * nilai idf)

TABEL 7
NILAI TD-IDF

tfidf_d1	tfidf_d2	tfidf_d3	tfidf_d4	tfidf_d5
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0,69897	0	0	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0
0	0	0,69897	0	0

5. *Menentukan Kata Identifikasi (Fitur)*

Identifikasi kemunculan kata dari hasil *idf* kemudian dihitung jumlah katanya pada setiap topik, kata yang memiliki nilai *tf-idf* tertinggi dari *tfidf_d1*, *tfidf_d2* dan seterusnya adalah kata yang mengidentifikasikan topik tersebut.

TABEL 8
IDENTIFIKASI KATA

d1	d3
penelitian	pemeriksaan
berupa	pajak
pengembangan	merupakan
sistem	serangkaian
pendukung	kegiatan
keputusan	mencari
spk	mengumpulkan
perencanaan	mengolah
kebutuhan	data
bahan	keterangan
baku	menguji
mempunyai	kepatuhan
batas	pemenuhan
masa	kewajiban
kadaluarsa	perpajakan
ketentuan	tujuan
diskon	rangka
pembelian	melaksanakan
jumlah	ketentuan
tertentu	peraturan
	perundang
	undangan
	perpajakan

6. *Hasil Perhitungan Klasifikasi*

TABEL 9
NILAI TD-IDF

Topik	Perhitungan
d1	0,333333
d2	0,291667
d3	0,3125
d4	0,291667
d5	0,291667

7. *Perhitungan Similarity Vector Space Model*
   a. Menghitung akar dari total *term keyword* di semua document dan akar dari term dari setiap document dari hasil *tokenizing*.

   Rumus: sqrt(jumlah term *keyword* atau jumlah *term document*)

   contoh :
   q (jumlah keyword) = 5
   d1(jumlah document 1) = 19
   d2(jumlah document 2) = 24

   q : sqrt(5) = 2,23606
   d1 : sqrt(19) = 4,35890
   d2 : sqrt(24) = 4,898979

   b. Setelah mendapatkan akar dari document dan *keyword* barulah menghitung *similarity*.

   Rumus : (jumlah *term keyword* di document * jumlah *term document*)/(akar dari *keyword* * akar dari *keyword*)
   Contoh :
   d1 = *keyword* muncul di document 1 sebanyak 4
   d2 = *keyword* muncul di document 2 sebanyak 1

   d1 : (4 * 19)/(2,23606*4,35890) = 7,79743
   d2 : (1 * 24)/(2,23606*4,89897) = 2,19089

Perhitungan pencarian *similarity* diatas antara *document* dan *keyword*, sehingga dapat mengurutkan *document* mana yang paling mirip dengan *keyword* pencarian. Dari hasil tersebut dapat dilihat bahwa *d1* memiliki hasil yang lebih besar. dengan hasil tersebut *d1* di atas peringkat pencarian dan *d2* dibawahnya.

III. HASIL DAN PEMBAHASAN

Penerapan model klasifikasi *Naïve Bayes* dan deteksi kesamaan dokumen dengan *Vector Space Model* diuji dengan beberapa makalah yang memiliki format file .pdf dan .docx. Berikut ini adalah uji coba yang dilakukan :

TABEL 11
HASIL UJI COBA

	1	2	3	4	5	6	7	8	9	10
Klasifikasi	5.53E-288	-	-	1.15E128	-	-	2.42E-206	7.28E254	-	8.54E-218
Similarity	91.52%	86.73%	76.54%	94.79%	-	62.61%	90.38%	90.84%	71.74%	96.52%

TABEL 10
DATA LATIH

Data Latih	
Dokumen	Topik Jurnal
Dokumen1.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen2.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen3.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen4.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen5.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen6.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen7.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen8.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen9.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen10.pdf	Ilmu Komputer dan Teknologi Informasi

Data pada tabel 10 adalah makalah yang menjadi data pelatihan. Makalah ini memiliki topik jurnal di bidang ilmu komputer dan teknologi informasi. Format file yang digunakan sebagai data pelatihan adalah .pdf.

TABEL 11
DATA UJI

Data Uji	
Dokumen	Topik Jurnal
Dokumen1.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen2.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen3.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen4.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen5.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen6.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen7.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen8.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen9.pdf	Ilmu Komputer dan Teknologi Informasi
Dokumen10.pdf	Ilmu Komputer dan Teknologi Informasi

Data dalam table 11 adalah makalah yang menjadi data uji, dimana makalah yang digunakan sebagai data uji sama dengan makalah yang digunakan pada data latih. Makalah ini memiliki topik jurnal di bidang Ilmu Komputer dan Teknologi Informasi. Format file yang digunakan sebagai data uji adalah .pdf.

IV. KESIMPULAN

Dalam beberapa dokumen data uji dengan format file .pdf, algoritma *Naïve Bayes* tidak dapat mengklasifikasikan ke dalam satu topik jurnal dengan tepat tetapi mengklasifikasikan menjadi beberapa topik jurnal sehingga mempengaruhi kinerja metoda *Vector Space Model.* Hasil perhitungan deteksi kesamaan teks oleh *Vector Space Model* dapat mencapai 90% ke atas. Hasil perhitungan deteksi kesamaan teks dengan metoda *Vector Space Model* juga sangat dipengaruhi oleh data

pelatihan. Semakin lengkap dan kompleks data pelatihan, maka semakin valid hasil pengujian kinerja *Vector Space Model*.

Saran untuk penelitian ini adalah pengujian kinerja algoritma Naïve Bayes dan metoda Vector Space Model dapat diuji pada dokumen teks dalam berbagai format file. Saran berikutnya adalah data latih dapat terintegrasi ke database neliti.com

REFERENSI

[1] "Neliti.com.
[2] M. W. Berry, J. Kogan., *Text Mining Applications and Theory*. Wiley, 2010.
[3] P. Manning, Christopher D, Raghavan "*Introduction to Information Retrieval*.California : Stanford University, 2008.
[4] M. Chenoweth, and M. Song "*Text Categorization in Encyclopedia of Data Warehouse & Data Mining*. IGI Global, 2009.
[5] A. Ryansyah and A.P.K. Dokumen, "Implementasi Algoritma TF_IDF pada Pengukuran Kesamaan Dokumen," No.1, pp. 1-10.
[6] T.M. Isa et al.,"Mengukur Tingkat Kesamaan Paragraf Menggunakan Vector Space Model untuk Mendeteksi Plagiarisme", 20013.
[7] Mushlihudin. Dan Zahrotun, L. 2017. "Perancangan Text Mining Pengelompokkan Penelitian Dosen Menggunakan Metode Shared Nearest Neighbor Dengan". P**rosiding SNATIF,** 849-855.
[8] *Setiawan, A., Astuti, I.F.,* dkk. "Klasifikasi dan Pencarian Buku Referensi Akademik Menggunakan Metoda Naïve Bayes Classifier (NBC) (Studi Kasus : Perpustakaan Daerah Provinsi Kalimanta Timur)", 10(1). 2015.
[9] Triana, A. "Pemanfaatan Metoda Vector Space Model dan Metoda Cosine Similarity pada Pemanfaatan Metoda Vector Space Model dan Metoda Cosine Similarity pada Fitur Deteksi Hama dan Penyakit Tanaman Padi", 2014.
[10] Wijanto, M. C., Teknik, S., dkk. "Sistem Pendeteksi Pengirim Tweet dengan Metoda Klasifikasi Naïve Bayes", 1, 172 – 182. 2015.